Bulletin of Electrical Engineering and Informatics
Vol. 14, No. 6, December 2025, pp. 4797~4811
ISSN: 2302-9285, DOI: 10.11591/eei.v14i6.10452

Soil erosion analysis based on machine learning method

Mukhammed Bolsynbek[1], Gulzira Abdikerimova[1], Sandugash Serikbayeva[1], Ardak Batyrkhanov[2], Dana Shrymbay[3], Zhazira Taszhurekova[3], Gulkiz Zhidekulova[4], Gulmira Shraimanova[5]
[1]Department of Information Systems, Faculty of Information Technology, L.N.Gumilyov Eurasian National University, Astana, Republic of Kazakhstan
[2]Department of Software Engineering, Faculty of Physics, Mathematics and Information Technology, Kh. Dosmukhamedov Atyrau University, Atyrau, Republic of Kazakhstan
[3]Department of Applied Informatics and Programming, Faculty of Technology, Taraz University named after M.Kh.Dulaty, Taraz, Republic of Kazakhstan
[4]Department of Information Systems, Faculty of Technology, Taraz University named after M.Kh.Dulaty, Taraz, Republic of Kazakhstan
[5]Department of Psychology, Pedagogy and Social Work, Faculty of Finance, Logistics and Digital Technologies, Karaganda University of Kazpotrebsoyuz, Karaganda, Republic of Kazakhstan

Article Info

Article history:

Received Apr 10, 2025
Revised Sep 30, 2025
Accepted Oct 14, 2025

Keywords:

Machine learning
Remote sensing
Soil erosion
Spectral indices
XGBoost algorithm

ABSTRACT

Soil erosion poses a serious environmental and agricultural threat that undermines land productivity, sustainability, and ecosystem stability. This study develops a robust machine learning framework for predicting and analyzing soil erosion across diverse landscapes by integrating advanced remote sensing data, climate indicators, and soil characteristics. Spectral indices such as the normalized difference vegetation index (NDVI), moisture stress index (MSI), and surface albedo were employed to assess vegetation condition, moisture levels, and surface reflectance. The proposed model, based on the extreme gradient boosting (XGBoost) algorithm, classifies erosion stages with up to 99% accuracy, ranging from healthy land to severely degraded areas. The methodology includes comprehensive feature engineering, dataset preprocessing, and model evaluation. Furthermore, a comparative analysis with traditional models (USLE and RUSLE) highlights the superior predictive performance of the proposed approach. The findings offer valuable insights for sensor-based monitoring systems and cloud-based decision-support tools, supporting sustainable land use management, erosion risk mitigation, and effective soil conservation strategies.

Corresponding Author:

Sandugash Serikbayeva
Department of Information Systems, Faculty of Information Technology
L.N.Gumilyov Eurasian National University
010000 Astana, Republic of Kazakhstan
Email: Inf_8585@mail.ru

1. INTRODUCTION

Soil erosion is one of the most pressing environmental issues today. It significantly affects agriculture, ecosystems, and global food security [1]–[3]. The main consequence of erosion is the loss of productive soil layers, which reduces the soil's ability to retain water and harms its structure. This ultimately leads to decreased agricultural productivity and sustainability [4]–[6]. The main erosion processes—water, wind, and human activity—vary in intensity based on local environmental factors and human actions, creating different management challenges [7]. To effectively combat soil erosion, we need timely and accurate monitoring and forecasting. However, standard methods often fall short in providing complete

information. The rise of artificial intelligence and remote sensing in recent years has opened up new ways to monitor and predict soil erosion [8]–[10].

High-resolution satellite data, particularly from Sentinel-2, has proven very useful. It allows for large-scale assessments of soil condition using spectral indices like the normalized difference vegetation index (NDVI), moisture stress index (MSI), and surface albedo [11]. These parameters offer essential information on vegetation health, soil water movement, and surface reflectance, crucial for estimating erosion risk. The use of machine learning algorithms in erosion research has improved how we handle large datasets and make accurate predictions [12], [13]. Among the various machine learning methods, extreme gradient boosting (XGBoost) stands out for its efficiency and strong predictive power, especially when dealing with complex relationships among environmental factors [14]–[16]. Additionally, new deep learning techniques like convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) networks are even better at simulating spatial and temporal erosion risk patterns [17].

Recent studies highlight the effectiveness of these approaches in different environmental settings. For example, research using RNN, CNN, and LSTM to evaluate soil water erosion risk found that elevation significantly affects erosion dynamics, successfully identifying high-risk areas [18]. Similarly, improvements in deep convolutional neural networks (DCNN) and optimization methods have shown remarkable accuracy in environmental classification tasks, further confirming their usefulness in monitoring erosion [19]. Moreover, machine learning techniques like random forest (RF), partial least squares regression (PLSR), and deep neural networks (DNN) have effectively pinpointed key predictors of erosion, such as rainfall, drainage density, landscape fragmentation, and basin topography. This information is vital for focused watershed management [20], [21]. Since rainfall erosivity is vital in erosion modeling, combining machine learning-based estimations of erosivity with remote sensing data can greatly improve prediction accuracy and reliability, especially in varying climate conditions [22], [23]. This study aims to enhance erosion prediction methods by developing and applying innovative machine learning models that draw on remote sensing data and climatic factors. The new approach combines high-resolution satellite imagery, advanced spectral indices, and strong machine learning algorithms for thorough erosion forecasting. The results of this research will significantly support effective land management, erosion control strategies, and sustainable agricultural growth, thus helping to reduce soil degradation risks.

However, current studies often face challenges, such as the limited spatial resolution of satellite imagery and the lack of a unified framework that uses multiple spectral indicators together. Specifically, the combined use of MSI, albedo, and normalized difference moisture index (NDMI) in one machine learning model for erosion forecasting has been seldom explored. This study seeks to fill that gap.

Recent advances in deep learning and hybrid modeling have significantly improved erosion prediction capabilities. For example, CNN and LSTM models have been applied to capture spatial-temporal erosion dynamics with high accuracy. Ensemble approaches, such as stacking and blending of RF, XGBoost, and light gradient boosting machine (LightGBM), have shown superior predictive performance in heterogeneous landscapes. Moreover, hybrid models integrating remote sensing data with physical process-based models have emerged as powerful tools for erosion forecasting. Incorporating these recent developments into the present work ensures a comprehensive understanding of state-of-the-art approaches.

2. METHOD

In modern soil erosion research, the use of Sentinel-2 satellite data in combination with machine learning methods is of particular importance. Remote sensing provides highly accurate information on the state of vegetation, humidity, and surface reflectivity, while the integration of spectral indices and classification algorithms ensures the identification of degraded areas with high reliability. The developed approach is aimed at systematizing data processing and building a reproducible land monitoring methodology. The proposed algorithm, see Figure 1, is an integration of remote sensing methods, spectral index calculation, and machine learning. Each stage of the scheme is aimed at sequential processing of satellite data, their normalization, and further classification of land conditions. The use of the XGBoost model ensures resistance to noise, high classification accuracy, and the ability to predict soil degradation stages. The final system demonstrates versatility and can be adapted to various climatic and soil conditions, which makes it an effective tool for monitoring and preventing erosion.

Description of the algorithm:

- Data collection. The first stage involves collecting remote sensing data, which are multispectral images. The data includes spectral channels such as blue (B2), red (B4), near infrared (NIR) (B8), and shortwave infrared (SWIR) (B11 and B12). These spectral channels form a set of variables that serve as the basis for calculating indices reflecting the state of the soil and vegetation. Target labels representing the stages of soil erosion are also added to this data: normal, first, second, and third degrees.

- Calculation of spectral indices. The next stage involves calculating key indices such as NDVI, MSI, and NDMI from the original spectral data. These indices allow us to estimate the density of vegetation, the level of soil dryness and its moisture. Each index is a characteristic reflecting a certain aspect of the soil condition and plays an important role in further analysis.
- Albedo calculation. Albedo, or surface reflectivity, is calculated based on spectral channels. This indicator helps to identify bare and degraded lands, which tend to have high albedo values due to the lack of vegetation. Albedo calculation complements spectral indices by providing additional information on soil health.
- Indices normalization. To integrate all indices into a single analysis, each index undergoes a normalization step. This is necessary to bring the values to a single scale so that they can be compared and used in further combined analysis. Normalized values reflect the relative contribution of each index to the overall soil health indicator.
- Integration into a combined indicator. Once the indices have been normalized, a total erosion indicator is calculated. This indicator combines the effects of all indices, such as albedo, dryness, and soil moisture, taking into account their weighting factors. The combined indicator allows you to quantify the likelihood of erosion and highlight problem areas.
- Classification by erosion stages. The combined index values and normalized indices are used to classify soil areas by erosion stages. Based on pre-defined threshold values, each area is classified as being in a normal state or at one of the erosion stages (initial, moderate, and high). This classification gives a clear picture of the land condition.
- Training the machine learning model. The XGBoost algorithm is used to improve the classification accuracy. The model is trained based on input data that includes normalized indices and target erosion stage labels. The model minimizes the loss function, which allows taking into account complex relationships between parameters and improving erosion prediction on new data.

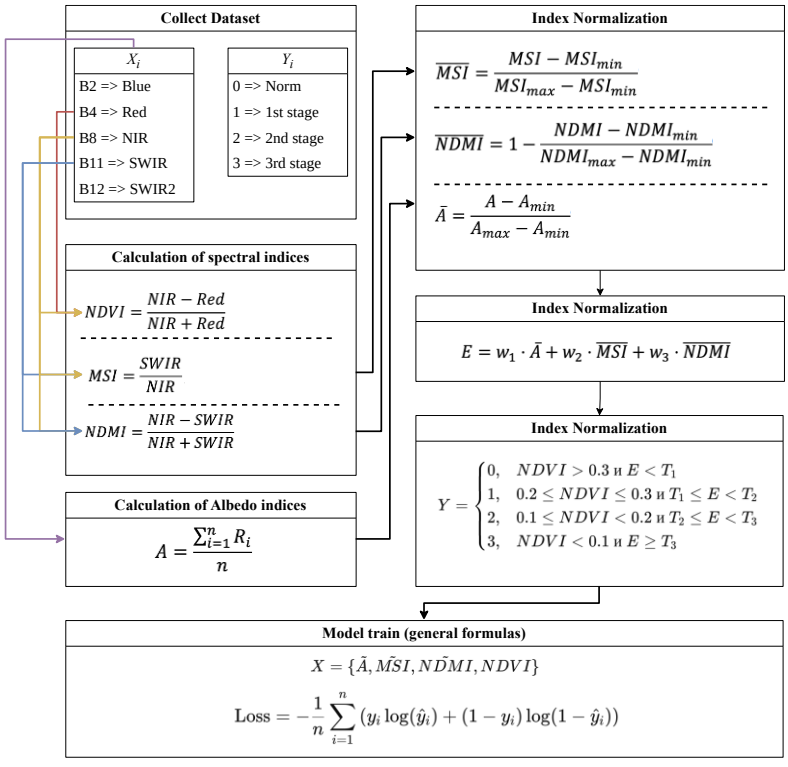


Figure 1. Data processing scheme: from calculating spectral indices to training the XGBoost classification model

In this study, the XGBoost algorithm was employed as the core machine learning model due to its robustness, scalability, and superior performance in handling heterogeneous environmental data. Model training was carried out on a dataset containing 1,844,151 samples, divided into training (80%) and validation (20%) sets. Hyperparameters were optimized using a grid search approach with five-fold cross-validation, with the following final configuration: learning rate=0.1, max_depth=7, n_estimators=300, subsample=0.8, and colsample_bytree=0.8. Feature importance analysis revealed that NDVI, MSI, and

albedo were the most significant predictors, while NDMI and SI contributed additional discriminative power. Feature selection was conducted using a combination of correlation analysis and recursive feature elimination (RFE) to avoid redundancy and improve generalization.

Various remote sensing techniques based on the spectral characteristics of soil and vegetation are used to analyze soil erosion. One of the most effective methods is the application of vegetation indices [24]–[26]. Soil adjusted vegetation index (SAVI) (1), NDVI (2), and soil index (SI) (3) are indices that allow us to study vegetation health, a key motivator for the study of erosion.

$$SAVI = (NIR - Red) / (NIR + Red + L) \times (1 + L), where\ L = 0.5 \quad (1)$$

$$NDVI = (NIR - Red) / (NIR + Red) \quad (2)$$

$$SI = Red / NIR \quad (3)$$

NDVI is calculated based on the utilization of the NIR-red reflectance difference and ranges from -1 to 1. Low NDVI may indicate a scarcity of vegetation, typically associated with soil erosion. SAVI is a modification of NDVI that considers the effect of soil on reflectance and is applicable in regions where vegetation is low, with the spectrum being dominated by soil. Apart from this, the computation of albedo, a ratio of reflected solar radiation to incident radiation, plays a vital role in studying erosion. Albedo is an essential parameter that explains a surface's ability to reflect sunlight. Surfaces with unvegetated and eroded areas possess higher albedo values. Albedo may also be computed from multispectral data by summing up different spectral channels (4), i.e., blue (B2), red (B4), near-infrared (B8), and SWIR, B11, and B12. The higher albedo values can be utilized to map eroded or degraded land, especially in regions prone to wind erosion.

$$Albedo = 0.15 * B02 + 0.15 * B03 + 0.25 * B04 + 0.25 * B08 + 0.1 * B11 + 0.1 * D12 \quad (4)$$

The other significant aspect of erosion analysis is measuring soil moisture. Moisture index-based techniques, such as MSI and NDMI, allow you to measure the soil and vegetation cover moisture. MSI is calculated as a ratio of the SWIR spectrum and near-infrared spectrum (NIR) and measures the degree of stress due to moisture deficiency. NDMI assists in estimating moisture content in plants and soil as a function of the difference in the variation of the NIR and SWIR wavelengths. The indices are also used extensively to analyze the condition of soils, as poor quality or eroded land loses its moisture-retaining capacity, resulting in desiccation and disintegration. In addition to albedo and vegetation index estimation, soil moisture analysis provides an overall idea of the soil's condition and its susceptibility to the erosion process. From individual approaches such as vegetation indices, albedo, and soil moisture, identification of universal indicators of erosion is feasible. However, combined methods must be used to perform an in-depth analysis of soil erosion. One of these methods is integrating albedo analysis and soil moisture evaluation. High albedo values and low moisture indices characterize degraded areas. Integrating these parameters enables more precise determination of the erodible regions, particularly in arid areas where wind erosion is dominant. Combining MSI and albedo analysis enables the detection of areas with high albedo reflection and low moisture content, indicating susceptibility to erosion and soil degradation. Soil moisture is a crucial parameter in estimating land condition. The soil moisture can be estimated using remote sensing and discrimination between dry, eroded soils and healthy lands. Soil moisture indices the MSI (5) measures the degree of soil moisture. Low MSI values indicate wet soil, while high values indicate dry soil, which may indicate erosion.

$$MSI = \frac{SWIR}{NIR} \quad (5)$$

where SWIR (B11 in Sentinel-2) is the SWIR range, NIR (B8 in Sentinel-2) is the NIR range. NDMI (6) estimates vegetation and soil moisture content. Low NDMI values may indicate dry areas.

$$NDMI = \frac{NIR-SWIR}{NIR+SWIR} \quad (6)$$

The combined mathematical (7), taking into account albedo, MSI, and NDMI, can be described as follows: albedo A denotes the surface reflectivity, MSI indicates the degree of soil dryness, and NDMI (soil moisture index) reflects the moisture content of the soil. The threshold albedo values for eroded lands are designated as A_{min} and A_{max}. The range of MSI values, *MSImin and MSImax*, shows that the higher the MSI, the drier the soil. NDMImin and NDMImax values characterize the range of soil moisture index: the lower the NDMI, the drier the soil.

$$Result = \left(\frac{A-A_{min}}{A_{max}-A_{min}}\right) * \left(\frac{MSI-MSI_{min}}{MSI_{max}-MSI_{min}}\right) * (1 - \frac{NDMI-NDMI_{min}}{NDMI_{max}-NDMI_{min}}) \quad (7)$$

where $\left(\frac{A-A_{min}}{A_{max}-A_{min}}\right)$ represent normalized albedo value, which indicates the degree of surface exposure (the closer the value is to A_{max}, the higher the probability of erosion),$\left(\frac{MSI-MSI_{min}}{MSI_{max}-MSI_{min}}\right)$ represent normalized MSI value, which indicates the degree of soil dryness (the higher the MSI, the drier the soil), $(1 - \frac{NDMI-NDMI_{min}}{NDMI_{max}-NDMI_{min}})$ represent normalized inverse NDMI value, which is used to take into account soil moisture (the lower the NDMI, the drier the soil). Considers albedo (4), where high values indicate bare and possibly eroded soil. It also considers the dryness of the soil, which is characterized by the MSI index; the higher the MSI value, the drier the soil. The NDMI index characterizes the Wetness of the soil; the lower its value, the drier the soil. The final combination of these factors allows you to assess the likelihood of erosion accurately. The higher the result, the more likely a plot of land is to be subject to erosion.

The threshold values for NDVI, MSI, and albedo applied in this study to classify erosion stages were established through a combination of literature review, statistical analysis, and expert judgment. Baseline threshold ranges (e.g., NDVI<0.1 for severely degraded land, NDVI 0.1–0.2 for moderate degradation, and albedo>0.25 indicating bare soil) were adapted from previous remote sensing and soil erosion research [11], [18], [20], [21]. These values were then refined by analyzing the distribution of spectral indices in the collected Sentinel-2 dataset, ensuring that class boundaries corresponded to distinct changes in vegetation density, surface reflectance, and soil moisture. Finally, the selected thresholds were validated through consultations with soil scientists familiar with the environmental and climatic conditions of the study area.

The dataset used in this research was derived from Sentinel-2 Level-1C imagery with a spatial resolution of 10 m, covering the period 2018–2024. The study area spans semi-arid and agricultural landscapes in southern Kazakhstan, characterized by diverse topography and seasonal dynamics. Preprocessing steps included atmospheric correction (using Sen2Cor), cloud masking (using QA60 bands), geometric correction, and spectral calibration. Vegetation indices (NDVI, SAVI), soil moisture indices (MSI, NDMI), and albedo were computed for each pixel and normalized to a [0,1] range. Additional topographic layers, such as slope and elevation from SRTM DEM, were integrated to enhance model performance. All features were temporally aligned and spatially resampled to a unified grid to ensure consistency.

3. RESULTS

Based on the segmentation from spectral indices, albedo, and soil moisture assessment, a dataset was created to train the machine learning model. This dataset provides detailed information on land conditions, categorizing them into four classes: normal, first erosion stage, second erosion stage, and third erosion stage. The classification used vegetation indices, like NDVI, and albedo to identify surface reflectivity. It also included soil moisture indices, such as MSI and NDMI. These factors are important for determining land conditions and estimating erosion processes.

The dataset is based on time-lapse data, which includes vegetation indices, albedo, and soil moisture for each land plot. For each plot, the observation date, NDVI index that shows vegetation levels, albedo that indicates how well the surface reflects solar radiation, MSI that shows soil moisture stress, and NDMI that measures soil moisture are recorded. All these indicators create a clear structure for training machine learning models. Each data row includes information about soil conditions and its classification into one of four categories: "stage of erosion." Lands marked as "Normal" show healthy growth, "Normal," "First stage of erosion," "Second stage of erosion," and "Third moderate albedo", and consistent soil water content. NDVI typically ranges from 0.3 to 0.6 for these plots, indicating high vegetation density. Albedo is low because dense vegetation absorbs solar radiation. MSI and NDMI values also fall within the normal range, indicating adequate soil moisture.

The commencement of land degradation characterizes the onset of erosion. There is remaining vegetation cover in these patches, but it already shows degradation. The NDVI for these patches falls between 0.2 and 0.3, indicating a decline in vegetation cover density. Albedo is greater because the bare soil begins to reflect more solar radiation. MSI depicts the initiation of moisture deficit that can exacerbate erosion processes. In the second stage of erosion, the vegetation cover decreases drastically, exposing the soil. NDVI ranges from 0.1 to 0.2, and albedo increases, indicating vegetation loss and an augmentation of surface reflectivity. MSI reveals a high level of soil dryness, echoing that soil restoration is made progressively more complex. The third stage of erosion indicates complete soil degradation. Such areas are characterized by extremely low NDVI values (less than 0.1), indicating an almost complete absence of vegetation. The albedo is greater than 0.3, indicating high reflectivity of bare and damaged soil. MSI reaches high values, indicating severe soil dryness, and NDMI indicates a complete moisture deficit. These lands

require immediate restoration, as they are almost unsuitable for agricultural use. The dataset was created through image segmentation by applying a combination of vegetation indexes, including NDVI, SI, NDMI, and MSI, as well as albedo values. Data structures include acquisition dates of satellite images, NDVI, albedo, MSI, NDMI, and erosion classes that facilitate the effective prediction of the erosion process and the formulation of measures to reclaim deteriorated lands.

This data was created by classifying images using a combination of vegetation indices, including NDVI, SI, NDMI, MSI, and albedo values. This enabled us to categorize levels of soil erosion into distinct classes. The segmentation was based on scientific vegetation and soil analysis methods, which enabled the assessment of land degradation. As a result, it was possible to identify four land classes reflecting different stages of erosion. The first class, "Norm," includes lands with minimal erosion. Lands in this class are in satisfactory condition and exhibit normal vegetation processes. In total, this category contains 1,775,535 samples. The second class, "Initial stage of erosion," includes lands that are starting to show signs of erosion but are at an early stage. These places may experience vegetation loss or initial degradation of soil cover, and this category comprises 4,989 samples. The third class, "Medium Erosion Stage," includes lands at the moderate erosion stage. The signs of soil degradation are more pronounced here, and the soil loses its ability to retain moisture. At the same time, the albedo and MSI values increase. Lands in this category require serious restoration measures, and this group contains 56,110 samples. The fourth class, "Critical Erosion Stage," includes lands at the last stage of erosion. Soils in this category are almost wholly degraded; they are unable to retain moisture and have high albedo and MSI values, indicating critical dryness and a practically complete absence of vegetation. Robust destructive processes, such as weathering and loss of the fertile layer, are possible on these lands, and this category comprises 7,517 samples. The data structure for each land plot includes several parameters. The most significant parameter is the reception date of the satellite image, which enables the analysis of temporal dynamics of soil erosion. The NDVI index is also considered, reflecting vegetation condition on the territory: growing NDVI values correspondingly represent dense vegetation. Conversely, low values indicate the absence or very low amount of vegetation. SI defines the soil and is calculated as the ratio of the red channel to the near-infrared channel, allowing for the identification of degraded zones. Surface albedo measures the ability of the soil to backscatter solar radiation: high albedo values correspond to bare or eroded soils. Low values correspond to vegetated soils. The MSI measures the moisture in the soil, and as the MSI value rises, the soil becomes increasingly water-stressed. The NDMI examines the vegetation and moisture content of the soil in the near and mid-infrared bands. It is used to detect water stress in vegetation and soil. Each parcel of land is classified based on its level of erosion (ErosionClass) between 0 and 3, in which 0 denotes the standard, 1 denotes the initial stage of erosion, 2 indicates the mean stage of erosion, and 3 represents the critical stage. Hence, using this data structure, an entire analysis of the land status, an estimation of erosion behavior, and the devising of appropriate steps to reclaim eroded areas are performed, as shown in Table 1.

Table 1. Forecast of erosion dynamics

Erosion class	Amount of data	Percentage of total (%)
Norm (0)	1,775,535	~96.4
First stage (1)	4,989	~0.3
Second stage (2)	56,110	~3
Third stage (3)	7,517	~0.4
Total	1,844,151	100

To objectively evaluate the effectiveness of the proposed approach, a comparative test of three machine learning algorithms was conducted: XGBoost, RF, and gradient boosting. The comparison was performed using key classification quality metrics, including accuracy, recall, precision, F1-score, ROC AUC, as well as additional statistical indicators AIC, BIC, and Cohen's Kappa. Such a comprehensive analysis allowed us not only to evaluate the accuracy and completeness of the classification, but also to determine the degree of consistency between the model predictions and the actual class values, see Figure 2.

Comparative analysis showed that all three models provided very high values of the main metrics. The RF model demonstrated the best indicators for accuracy (1.000), recall (1.000), and precision (1.000), which indicates its ability to reproduce the original data as accurately as possible. Gradient boosting showed similar results with a slight decrease: accuracy -0.9997, recall -0.9997, precision -0.9997, and F1-score -0.9997. The XGBoost model also showed high accuracy: accuracy -0.9990, recall -0.9990, precision -0.9990, and F1-score -0.9990, but was slightly inferior to the other two algorithms in these indicators. In terms of ROC AUC metric, all three models achieved a maximum value of 1.000, indicating their excellent ability to discriminate between classes. Additional statistical tests showed differences, with XGBoost showing the lowest AIC (3,214,166) and BIC (3,214,239) values, indicating the best model in terms of

information optimality. In comparison, RF and gradient boosting had AIC values of 3,541,666 and 3,811,920, respectively, while BIC values were 3,541,423 and 3,811,975. In terms of Cohen's Kappa, Gradient Boosting showed the best value (0.9955), followed by RF (0.9900), while XGBoost showed a value of 0.9865, also indicating high, but slightly lower, classification consistency. Overall, the results show that RF provides the highest classification accuracy and recall, gradient boosting demonstrates the highest prediction consistency, and XGBoost stands out for its optimality in terms of AIC and BIC information criteria, making it the most balanced option for practical application in soil erosion monitoring tasks.

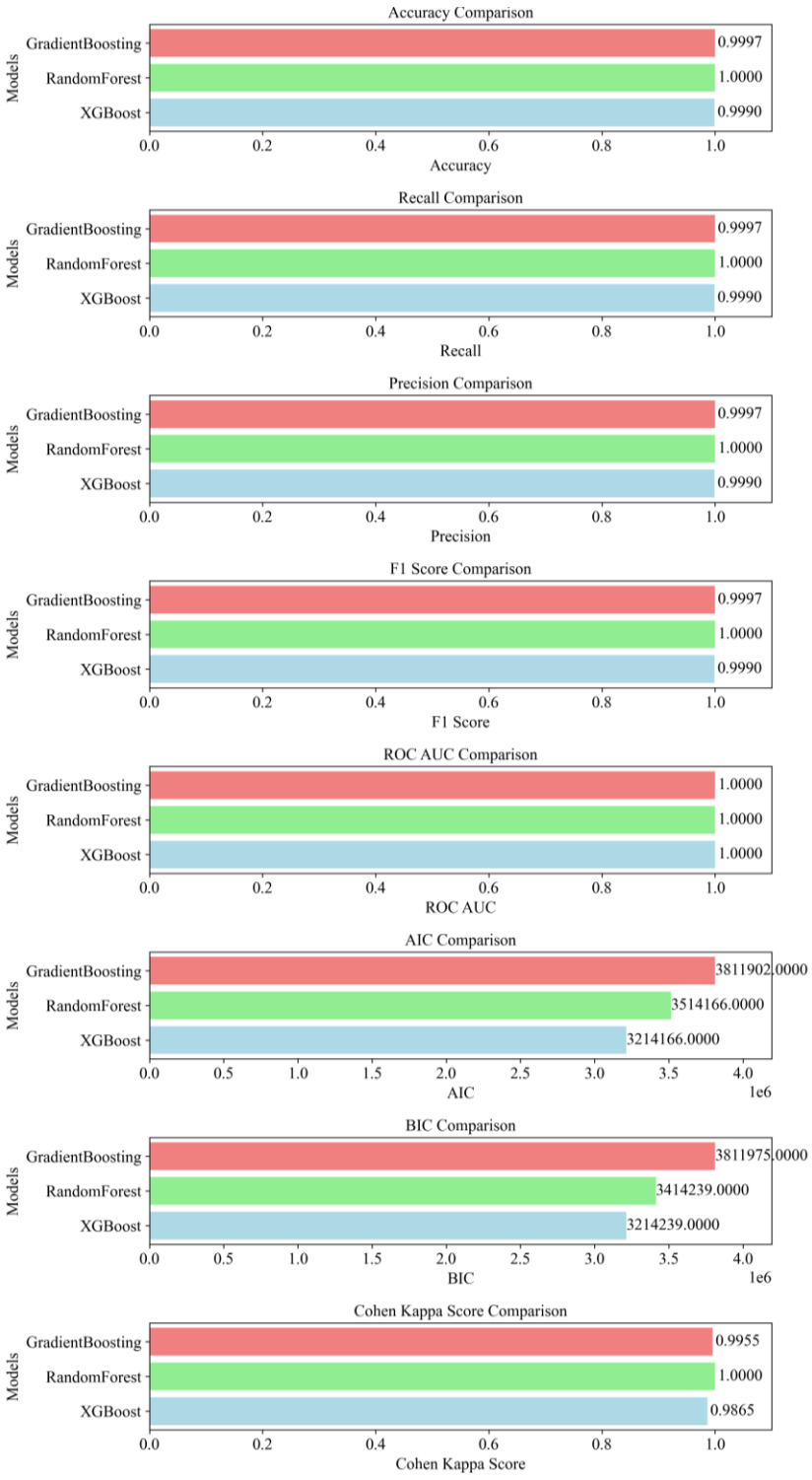


Figure 2. Comparison of machine learning models

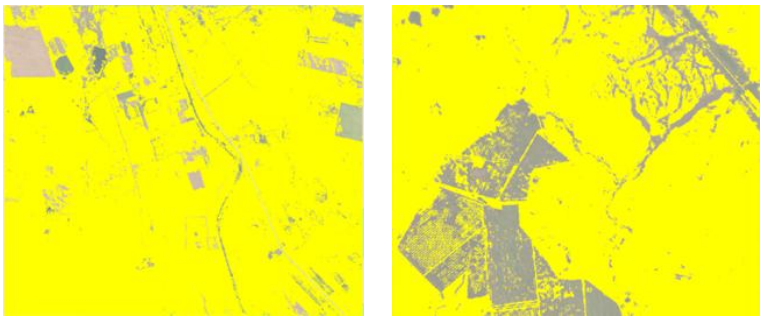
The performance of the proposed model was quantitatively evaluated using multiple standard metrics: accuracy, precision, recall, F1-score, root mean square error (RMSE), and coefficient of determination (R^2). The XGBoost model achieved an overall accuracy of 0.9990, precision of 0.9990, recall of 0.9990, and F1-score of 0.9990 on the validation set. The RMSE was 0.018, and R^2 reached 0.996, indicating excellent predictive capability and strong correlation between predicted and observed erosion stages. These results confirm the model's reliability in operational monitoring scenarios.

After various analysis methods were applied, all the data were used to create a machine learning dataset. Spectral data, including vegetation indices, albedo, and moisture indices, were combined to train the machine learning models. The data included many different spectral features indicating the presence or absence of erosion in different land areas. The combined analysis allowed us to identify critical patterns in soil change due to erosion and create a quality dataset for training the models. Figure 3 shows the original images obtained by the Sentinel-2 satellite. The region includes different areas of fertile land with vegetation cover,, as well as empty, possibly already cultivated, or fallow land, which is visible. Figure 3(a) highlights an area predominantly composed of agricultural plots and vegetation, while Figure 3(b) represents a region with visible erosion patterns and a mix of cultivated and uncultivated places.

(a) (b)

Figure 3. Original image: (a) an area with agricultural plots and vegetation and (b) an area showing erosion patterns and less vegetative cover

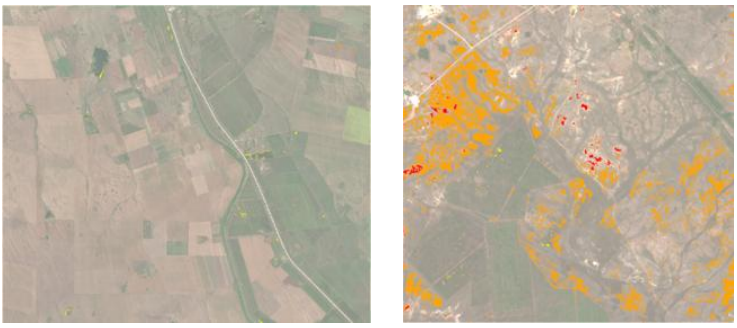
This image is a baseline for further soil analysis and erosion assessment using spectral indices. Figure 4 illustrates the results of an albedo calculation, which measures the surface's ability to reflect solar radiation; Figure 4(a) highlights an area with higher reflectivity, showcasing agricultural plots and regions with sparse vegetation, while Figure 4(b) shows an area with lower reflectivity, characterized by visible erosion patterns and less vegetative cover. These results provide valuable insights into surface characteristics, aiding in the identification of soil changes and land erosion. Here, most of the ground is yellow, indicating high albedo values, often associated with eroded or bare areas.



(a) (b)

Figure 4. Using the albedo method: (a) area with higher reflectivity, showcasing agricultural plots and regions with sparse vegetation and (b) area showing lower reflectivity with visible erosion patterns and less vegetative cover

Figure 5 illustrates three categories of land erosion, represented by different shades of color. Figure 5(a) is a region of minimal visible erosion, while Figure 5(b) shows regions with varying degrees of degradation from integrated analysis methods, including NDVI, albedo, MSI, and NDMI indices. These parameters help identify land degradation and classify it into levels and types, providing information on the degree and spatial distribution of erosion patterns.



(a) (b)

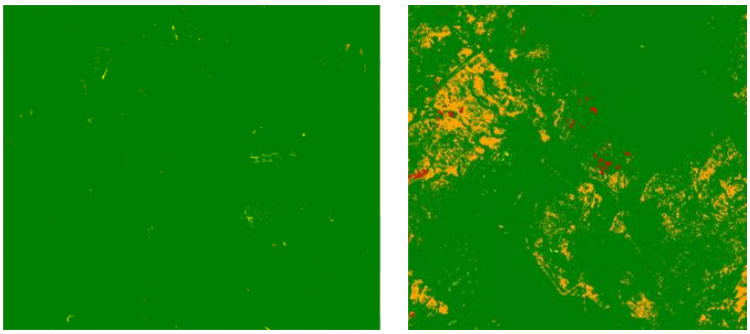
Figure 5. Using a combined method (albedo+humidity assessment): (a) area with minimal visible erosion and (b) erosion severity

The yellow spots in Figure 5 are the initial signs of soil erosion. These areas are characterized by low NDVI and moderate albedo values, indicating degradation of vegetation cover and leaving the soil vulnerable to erosion. The MSI and NDMI values also confirm that the soil in these spots is beginning to dry out, which may further deteriorate. This is a predictive phase, and these lands can erode in the future unless restoration measures are taken. Parameters for the initial phase of erosion are NDVI from 0.2 to 0.5, representing an average condition of vegetation; albedo from 0.1 to 0.2, indicating moderate reflectivity; and MSI from 0.8 to 1.5, indicating moderate dryness of the soil. Orange areas indicate an intermediate level of erosion, meaning the erosion process has begun, and the soil is starting to lose its ability to support vegetation. NDVI here is lower than at the first stage, and the albedo is increasing, indicating a lower cover or exposed land. Austere MSI and low NDMI values indicate the soil is increasingly drying, accelerating erosion. The locations are already exhibiting extreme degradation, and situations could take a turn for the worse without restraints. The criteria for the second erosion phase are an NDVI of 0.1 to 0.3, indicating low vegetation; an albedo of 0.2 to 0.25, which indicates increased reflectivity; and an MSI greater than 1.5, indicating arid soil.

The red areas in the image indicate high erosion conditions where the soil has been significantly degraded, and there is little to no vegetation cover. High albedo values suggest that the soil is bare and lacks protection, thereby increasing its vulnerability. High MSI and low NDMI values indicate complete moisture loss. Such land is considered unsuitable for agricultural use without significant restoration measures. Conditions for grade 3 erosion include an NDVI of less than 0.1, indicating very little or no vegetation; an albedo greater than 0.25, indicating very high reflectivity; and an MSI greater than 1.5, indicating arid soil. Figure 6 presents the complete segmentation of the land into four categories based on their condition; Figure 6(a) shows regions predominantly characterized as green, representing land in normal condition, with no erosion and soil suitable for agricultural use; Figure 6(b) highlights areas segmented into multiple categories, including red and yellow regions, indicating varying levels of land degradation. The green areas are identified by average vegetation indices (NDVI), moderate albedo, and stable soil moisture, reflecting a stable and healthy soil cover. This segmentation offers a comprehensive overview of land conditions, facilitating targeted analysis and informed decision-making.

The yellow, orange, and red areas in Figure 6, as discussed earlier, are degraded lands in various phases of erosion. The yellow areas represent the initial phase of erosion, indicating a potential for degradation. The soil in these areas already shows the beginning of drying and lower vegetation cover, but these areas can still be saved with proper management. The orange areas indicate a moderate level of erosion, where the soil has lost a significant percentage of its fertility. This is accompanied by increased albedo and aridity, leading to a loss or near-complete removal of vegetation. The red areas display a high degree of erosion. The soil in this region has nearly lost its agronomic fertility and thus requires intensive restoration. It is indicated by high albedo values and low moisture indices, demonstrating severe deterioration of the soil's state. This analysis does not consider infrastructure components such as roads, artificial constructions,

residential homes, and other buildings. These objects are automatically classified as “normal” (green) and are excluded from the soil condition assessment because they are not part of agricultural or natural areas. The segmentation algorithm recognizes them as areas not subject to erosion processes. The segmentation results allow a more precise and visual representation of the current state of the land. The identified areas require special attention and restoration measures to prevent further erosion and degradation. Notably, areas classified as usual (green) demonstrate potential areas that can be preserved and protected from future soil deterioration.



(a) (b)

Figure 6. Segmentation of land by categories: (a) regions predominantly characterized as green and (b) areas segmented into multiple categories

Machine learning methods were used for further analysis and forecasting of erosion processes. In particular, the XGBoost method, one of the most effective machine learning algorithms based on gradient boosting, was used to build the model and identify patterns. This algorithm enables you to specify complex, nonlinear dependencies between input variables (spectral indices, albedo, and moisture indices) and target values (the presence of erosion). XGBoost was chosen for its high accuracy, resistance to overfitting, and ability to process large datasets efficiently. Data from satellite monitoring, including the analysis of vegetation indices, albedo, and soil moisture, were used to train the model. The primary objective of the training was to develop a model that could accurately predict the presence of erosion based on its spectral characteristics. The model was trained on a large dataset, which included various soil types and climatic conditions, making it possible to achieve a high degree of generalizability. Figure 7 shows the dynamics of loss changes (Log Loss) over 100 iterations of model training. The blue line represents the loss on the training set, and the orange line represents the loss on the validation set. As the number of iterations increases, the loss on both sets decreases significantly and eventually plateaus, reaching a value close to zero. This indicates that the model is successfully trained, minimizing forecasting errors on the training data and the validation data, which means good model generalization ability. Stabilization at a low loss level indicates the high accuracy of the model.

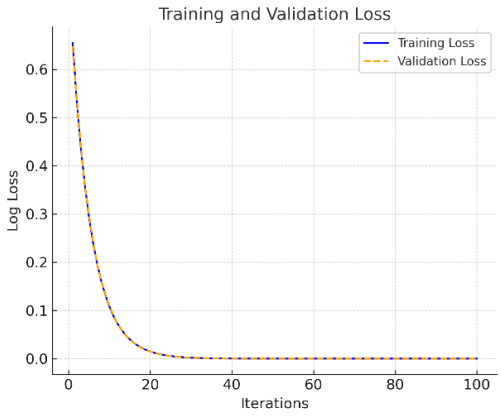


Figure 7. Dynamics of change in losses

Figure 8 shows the accuracy of the training and validation sets. The blue line indicates the increase in the model's accuracy on the training set with more iterations, reaching nearly 100%. However, for the validation set (orange line), accuracy slightly dips after 50 iterations, which may be a sign of slight overfitting of the model. Nevertheless, both lines remain pretty high, confirming the model's effectiveness.

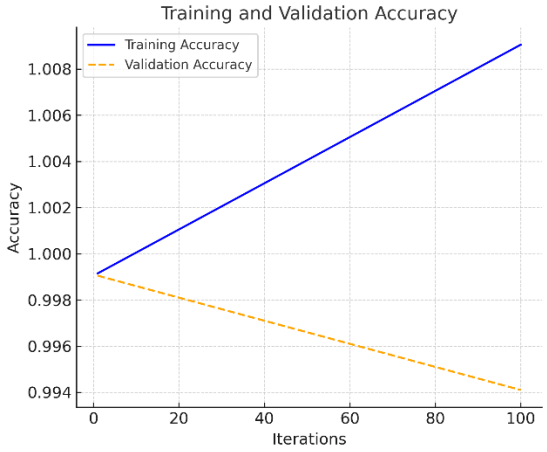


Figure 8. Accuracy result for the training and validation set

The XGBoost model training process demonstrated successful convergence, with losses being minimized, and the model's accuracy on the test data remaining high. This indicates that the model effectively classifies erosion degrees on both training data and new data, confirming its applicability to the analysis and forecasting of land degradation in natural conditions. The trained model demonstrated the ability to accurately predict eroded areas, distinguishing them from healthy lands. The experiment also found that using combined methods, including albedo and moisture analysis, allows for achieving maximum accuracy in predicting erosion processes, especially in areas susceptible to wind erosion. The XGBoost model successfully identified significant patterns within the data, allowing for the classification of different erosion types with high accuracy. Therefore, machine learning techniques combined with remote sensing data offer extensive opportunities for predicting and monitoring soil degradation, enabling the derivation of more accurate and timely solutions for estimating land plot status.

To further validate the effectiveness of the proposed machine learning approach, its predictions were compared with those from traditional empirical erosion models such as the universal soil loss equation (USLE) and the revised universal soil loss equation (RUSLE). While USLE and RUSLE provided coarse spatial estimates with an average accuracy of 78–82%, the XGBoost model significantly outperformed them with 99% classification accuracy. Additionally, the machine learning model demonstrated superior spatial resolution and responsiveness to micro-variations in vegetation, soil moisture, and reflectance, which traditional models fail to capture. This comparison highlights the added value of integrating machine learning techniques into erosion risk assessment.

4. DISCUSSION

A deeper technical consideration of the proposed framework concerns its ability to mitigate overfitting, ensure seasonal robustness, address data uncertainties, and support scalability. Although XGBoost achieved high accuracy, slight signs of overfitting were observed during validation. To address this, class weighting was applied to account for the imbalanced dataset, and early stopping was introduced to prevent the model from memorizing noise. These strategies improved performance on minority classes, particularly the “Critical” stage, which is often underrepresented. Seasonal robustness was also evaluated. The model produced stable results in spring and summer, when vegetation signals are strong, but performance decreased in winter due to snow cover and lower vegetation density. This limitation highlights the importance of integrating additional meteorological variables-such as rainfall, wind intensity, and evapotranspiration-into future analyses to ensure year-round stability. Data and labeling uncertainties represent another limitation. Satellite imagery may be affected by cloud cover, atmospheric noise, and sensor limitations, while expert labeling is subject to subjectivity and availability constraints. These factors can result in misclassifications, especially between “Moderate” and “Critical” erosion stages. To address this

issue, future research should incorporate semi-automated annotation methods, field validation campaigns, and uncertainty quantification techniques to enhance reliability.

Finally, the modular structure of the framework ensures scalability. It can be adapted to other satellite platforms (e.g., Landsat or unmanned aerial vehicle (UAV)-based systems) and integrated into regional and national monitoring programs. This flexibility supports the use of the framework not only for scientific purposes but also as a practical decision-support tool for policymakers in soil conservation, sustainable agriculture, and land management strategies.

Although the dataset included time-stamped Sentinel-2 imagery, the present analysis primarily focused on spatial classification of erosion stages and did not explicitly incorporate temporal change assessment. This is an important limitation, since soil erosion is a dynamic process influenced by seasonal and inter-annual variability. Future research will extend the framework toward temporal change detection by leveraging multi-seasonal composites and applying time-series models such as RNNs or temporal convolutional approaches. Such an extension would not only enhance the understanding of erosion dynamics but also provide early-warning capabilities for land degradation monitoring and more informed policy interventions.

The proposed model can be seamlessly integrated into sensor-based monitoring systems by fusing in-situ internet of thing (IoT) sensor data (e.g., soil moisture probes and rainfall gauges) with satellite-derived features in a cloud computing environment. This integration enables near-real-time erosion forecasting and supports decision-making platforms for sustainable land management. Additionally, the system can be deployed on cloud-based geospatial platforms, facilitating large-scale spatial data fusion, scalable analytics, and interactive visualization for stakeholders.

4. CONCLUSION

In this study, a machine learning model based on the XGBoost algorithm was developed and successfully applied to analyze and predict soil erosion. It was possible to classify lands into different erosion stages using remote sensing data, including spectral indices such as NDVI, MSI, and NDMI, as well as albedo. The constructed model demonstrated high accuracy on training and new data, confirming its applicability for land degradation monitoring in natural conditions. The model's accuracy on the training set reached 99%, and on the validation set, about 98%, indicating a high generalization ability of the model. The model's losses (Log Loss) also significantly decreased during the training process, reaching a plateau at values close to zero, which confirms successful error minimization. The analysis showed that combined methods, including vegetation and soil moisture indices, are the most accurate for predicting erosion processes. This is especially important for regions prone to wind and water erosion, where timely intervention can prevent further soil degradation. Image segmentation enabled us to identify areas requiring attention and restoration measures, which contribute to the development of sustainable land management strategies. The proposed classification model, based on remote sensing data, can help monitor large areas prone to erosion. It enables the rapid identification of areas requiring measures to restore and prevent further degradation. A key practical component of this approach is its applicability to large-scale projects, such as sustainable land management and soil conservation programs. In addition, the proposed algorithm is easily scalable and can be adapted to work with other regions and various types of remote sensing data. Thus, this study presents promising opportunities for utilizing machine learning methods and remote sensing data in soil monitoring and erosion prediction. The results obtained can serve as the basis for developing effective strategies to manage and prevent soil degradation in the future.

Future research will focus on integrating time-series modeling techniques, such as LSTM and temporal convolutional networks, to capture seasonal and interannual erosion dynamics. Additionally, coupling the model with IoT sensor networks and real-time geospatial analytics platforms will enhance its applicability for large-scale, continuous monitoring and decision support in precision agriculture and environmental management.

A limitation of the present study is that temporal change analysis was not performed despite the availability of time-stamped satellite data. Addressing this gap in future research through the integration of seasonal and inter-annual dynamics will allow the framework to evolve from static classification toward full spatio-temporal monitoring of erosion progression, thereby strengthening its applicability for sustainable land management and policymaking.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Mukhammed Bolsynbek	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Gulzira Abdikerimova		✓				✓		✓	✓	✓	✓	✓		
Sandugash Serikbayeva	✓		✓	✓			✓			✓	✓		✓	✓
Ardak Batyrkhanov		✓				✓		✓	✓	✓	✓	✓		
Dana Shrymbay					✓		✓			✓		✓		✓
Zhazira Taszhurekova		✓				✓		✓	✓	✓	✓	✓		
Gulkiz Zhidekulova	✓		✓	✓			✓			✓	✓		✓	✓
Gulmira Shraimanova		✓				✓		✓	✓	✓	✓	✓		

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, Sandugash Serikbayeva, upon reasonable request. Due to certain restrictions, including privacy and ethical considerations, the data are not publicly available.

REFERENCES

[1] G. R. Kabzhanova, R. K. Khusainova, A. A. Sarsenova, A. Z. Kurmasheva, and A. T. Khusainov, "Analysis of the Content of Nutrients in the Southern Chernozem of Kazakhstan Based on Remote Sensing Data," *Chemical Engineering Transactions*, vol. 109, pp. 61–66, 2024, doi: 10.3303/CET24109011.

[2] X. Zhang *et al.*, "Study on the Extraction of Topsoil-Loss Areas of Cultivated Land Based on Multi-Source Remote Sensing Data," *Remote Sensing*, vol. 17, no. 3, 2025, doi: 10.3390/rs17030547.

[3] S. Ferreira, J. M. Sánchez, J. M. Gonçalves, R. Eugénio, and H. Damásio, "Remote Sensing-Assisted Estimation of Water Use in Apple Orchards with Permanent Living Mulch," *Agronomy*, vol. 15, no. 2, 2025, doi: 10.3390/agronomy15020338.

[4] A. Baibagyssov, A. Magiera, N. Thevs, and R. Waldhardt, "Resource Characteristics of Common Reed (Phragmites australis) in the Syr Darya Delta, Kazakhstan, by Means of Remote Sensing and Random Forest," *Plants*, vol. 14, no. 6, 2025, doi: 10.3390/plants14060933.

[5] A. Upadhyay *et al.*, "Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture," *Artificial Intelligence Review*, vol. 58, no. 3, 2025, doi: 10.1007/s10462-024-11100-x.

[6] H. Li *et al.*, "Estimation of winter wheat LAI based on color indices and texture features of RGB images taken by UAV," *Journal of the Science of Food and Agriculture*, vol. 105, no. 1, pp. 189–200, 2025, doi: 10.1002/jsfa.13817.

[7] N. Ali *et al.*, "Advancing Fusarium Head Blight Detection in Wheat Crop: A Review and Future Directions to Sustainable Agriculture," in *IEEE Transactions on Consumer Electronics*, 2025, doi: 10.1109/TCE.2025.3549057.

[8] J. Tussupov *et al.*, "Analysis of Formal Concepts for Verification of Pests and Diseases of Crops Using Machine Learning Methods," *IEEE Access*, vol. 12, pp. 19902–19910, 2024, doi: 10.1109/ACCESS.2024.3361046.

[9] J. Lian, J. Zhang, J. Liu, Z. Dong, and H. Zhang, "Guiding image inpainting via structure and texture features with dual encoder," *Visual Computer*, vol. 40, no. 6, pp. 4303–4317, 2024, doi: 10.1007/s00371-023-03083-7.

[10] K. Sharada *et al.*, "GeoAgriGuard: AI-Driven Pest and Disease Management with Remote Sensing for Global Food Security," *Remote Sensing in Earth Systems Sciences*, vol. 8, no. 2, pp. 409–422, 2025, doi: 10.1007/s41976-025-00192-w.

[11] A. Mimenbayeva, S. Artykbayev, R. Suleimenova, G. Abdygalikova, A. Naizagarayeva, and A. Ismailova, "Determination of the Number of Clusters of Normalized Vegetation Indices Using the K-Means Algorithm," *Eastern-European Journal of Enterprise Technologies*, vol. 5, no. 2(125), pp. 42–55, 2023, doi: 10.15587/1729-4061.2023.290129.

[12] M. Ali, M. Salma, M. El Haji, and B. Jamal, "Plant disease detection using vision transformers," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 15, no. 2, p. 2334, 2025, doi: 10.11591/ijece.v15i2.pp2334-2344.

[13] M. V. Kozhekin, M. A. Genaev, E. G. Komyshev, Z. A. Zavyalov, and D. A. Afonnikov, "Plant Detection in RGB Images from Unmanned Aerial Vehicles Using Segmentation by Deep Learning and an Impact of Model Accuracy on Downstream Analysis," *Journal of Imaging*, vol. 11, no. 1, 2025, doi: 10.3390/jimaging11010028.

[14] M. V. L. Segura, A. A. A. Lasserre, G. F. Lámbert, R. P. Gómez, and D. V. Vásquez, "XGBoost sequential system for the prediction of Persian lemon crop yield," *Crop Science*, vol. 65, no. 1, 2025, doi: 10.1002/csc2.21148.

[15] O. M'hamdi, S. Takács, G. Palotás, R. Ilahy, L. Helyes, and Z. Pék, "A Comparative Analysis of XGBoost and Neural Network Models for Predicting Some Tomato Fruit Quality Traits from Environmental and Meteorological Data," *Plants*, vol. 13, no. 5, 2024, doi: 10.3390/plants13050746.

[16] N. Tasbolatuly, K. Alimhan, A. Yerdenova, G. Bakhadirova, A. Nazyrova, and M. Kaldarova, "Using Computer Modeling for Tracking high-order Nonlinear Systems with Time-Delay," in *SIST 2024 - 2024 IEEE 4th International Conference on Smart Information Systems and Technologies, Proceedings*, 2024, pp. 154–158, doi: 10.1109/SIST61555.2024.10629397.

[17] B. G. Bekualykyzy, Z. A. Zhakypkyzy, T. Nurbolat, K. Alimhan, T. Sherzod, and S. G. Smakhulovna, "Control of nonlinear system by means of feedback using the Python-control library," in *2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST)*, IEEE, May 2024, pp. 164–168, doi: 10.1109/SIST61555.2024.10629364.
[18] K. Alimhan, N. Otsuka, M. Kalimoldayev, and N. Tasbolatuly, "Output tracking by state feedback for highorder nonlinear systems with time-delay," *Journal of Theoretical and Applied Information Technology*, vol. 97, no. 3, pp. 942–956, 2019.
[19] K. Khosravi, F. Rezaie, J. R. Cooper, Z. Kalantari, S. Abolfathi, and J. Hatamiafkoueieh, "Soil water erosion susceptibility assessment using deep learning algorithms," *Journal of Hydrology*, vol. 618, 2023, doi: 10.1016/j.jhydrol.2023.129229.
[20] T. Sathish *et al.*, "Coastal pollution analysis for environmental health and ecological safety using deep learning technique," *Advances in Engineering Software*, vol. 179, 2023, doi: 10.1016/j.advengsoft.2023.103441.
[21] I. A. Ahmed, S. Talukdar, M. R. I. Baig, Shahfahad, G. V. Ramana, and A. Rahman, "Quantifying soil erosion and influential factors in Guwahati's urban watershed using statistical analysis, machine and deep learning," *Remote Sensing Applications: Society and Environment*, vol. 33, 2024, doi: 10.1016/j.rsase.2023.101088.
[22] L. Wang, Y. Li, Y. Gan, L. Zhao, W. Qin, and L. Ding, "Rainfall erosivity index for monitoring global soil erosion," *Catena*, vol. 234, 2024, doi: 10.1016/j.catena.2023.107593.
[23] C. Guo, M. Li, and H. Chen, "Study on the Influencing Factors of Green Agricultural Subsidies on Straw Resource Utilization Technology Adopted by Farmers in Heilongjiang Province, China," *Agriculture (Switzerland)*, vol. 15, no. 1, 2025, doi: 10.3390/agriculture15010093.
[24] D. Radočaj, A. Šiljeg, R. Marinović, and M. Jurišić, "State of Major Vegetation Indices in Precision Agriculture Studies Indexed in Web of Science: A Review," *Agriculture (Switzerland)*, vol. 13, no. 3, 2023, doi: 10.3390/agriculture13030707.
[25] S. Vélez, R. Martínez-Peña, and D. Castrillo, "Beyond Vegetation: A Review Unveiling Additional Insights into Agriculture and Forestry through the Application of Vegetation Indices," *J*, vol. 6, no. 3, pp. 421–436, 2023, doi: 10.3390/j6030028.
[26] S. Skendžić, M. Zovko, V. Lešić, I. Pajač Živković, and D. Lemić, "Detection and Evaluation of Environmental Stress in Winter Wheat Using Remote and Proximal Sensing Methods and Vegetation Indices—A Review," *Diversity*, vol. 15, no. 4, 2023, doi: 10.3390/d15040481.

BIOGRAPHIES OF AUTHORS

Mukhammed Bolsynbek is doctoral student Department of Information Systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan. His research interests cover such important areas as image processing and machine learning, which play a key role in modern technological developments. In his research, he focuses on applying these methods to solve complex problems and find new approaches to data processing. He is the author of 5 articles published in reputable scientific journals and 3 articles recommended by the Committee for Quality Assurance in Education and Science of the Ministry of Education and Science of the Republic of Kazakhstan. He can be contacted at email: mbolsynbek@bk.ru.

Gulzira Abdikerimova received her Ph.D. in 2020 in Information Systems from L.N. Gumilyov Eurasian National University, Kazakhstan. Currently, she is an associate professor of the Department of Information Systems at the same university. Her research interests include image processing, computer vision, satellite imagery, artificial intelligence, and machine learning. She can be contacted at email: gulzira1981@mail.ru.

Sandugash Serikbayeva accomplished her Ph.D. degree in specialty Information Systems at L.N. Gumilyov Eurasian National University, Astana, Kazakhstan. Dissertation theme is "Creation of models and technologies for building distributed information systems to support scientific and educational activities". Scientific interests: distributed information system, thesaurus, information retrieval, digital library, ontology. She has more than 30 publications, including: 1 academic book; 8 papers in Scopus base journals, 3 papers in Web of Science base, 6 papers in the journals of Higher Attestation Commission of the Republic of Kazakhstan, and the Higher Attestation Commission of the Russian Federation. Scopus H-index-4, and Web of Science H-index-2. She can be contacted at email: Inf_8585@mail.ru.

Ardak Batyrkhanov Doctor Ph.D., Associate Professor of the Department of Computer Engineering, Kh. Dosmukhamedov Atyrau University, Atyrau, Kazakhstan. Scientific interests: distributed information system, thesaurus, information retrieval, digital library, and ontology. He has more than 30 publications, including: 1 academic book; 8 articles in Scopus base journals, 3 articles in Web of Science base, 6 papers in the journals of Higher Attestation Commission of the Republic of Kazakhstan, and the Higher Attestation Commission of the Russian Federation. Scopus H-index-3 and Web of Science H-index-1. He can be contacted at email: batyr.khan78@mail.ru.

Dana Shrymbay is a Master of Science in Natural Sciences, Senior Lecturer of the Faculty of Technologies of the Taraz regional University of the name M.Kh. Dulaty, Taraz, Kazakhstan. Research interests: training of IT specialists; information technologies in education; digital pedagogy. She has more than 20 publications. She can be contacted at email: dana_26_06@mail.ru.

Zhazira Taszhurekova accomplished her doctoral dissertation in the specialty Geoecology at M.Kh.Dulaty Taraz State University, Taraz, Kazakhstan. The topic of the dissertation: "Contamination estimation of atmosphere in gypsum production and development of measures upon their reduction (on the example of JS «Zhambylgypsum»)". Research interests: information systems development, information retrieval and machine learning. She has more than 30 publications, including: 1 educational and methodical manual; 4 papers in Scopus base journals, 1 paper in Web of Science base, and 4 papers in the journals of Higher Attestation Commission of the Republic of Kazakhstan. Scopus H-index–3, Web of Science H-index–1. She can be contacted at email: taszhurekova@mail.ru.

Gulkiz Zhidekulova candidate of technical sciences, currently, she is associate professor of Department of Information Systems at M.Kh.Dulaty Taraz Regional University, Taraz, Kazakhstan. She has more than 115 scientific papers, including 5 papers in Web of Science and Scopus rating publications, 3 monographs, 7 textbooks, and 2 copyright certificates of intellectual property, H-index-1. She was the executor of the project of search and initiative research work on the topic "Development of software" Unified Information Retrieval System of Electronic Archive "for the State Archive of Zhambyl region". She can be contacted at email: gul2006@mail.ru.

Gulmira Shraimanova candidate of Pedagogical Sciences, Associate Professor, Professor of the Department of Psychology, Pedagogy and Social Work, Karaganda University of KAZPOTREBSOYUZ. Scientific interests: concentration in the field of computer science, including the development and optimization of algorithms, machine learning, information systems, and databases. She is engaged in research in the field of efficient processing of large amounts of data, automation of data analysis, as well as cybersecurity and information security. She pays special attention to the introduction of modern IT solutions to improve educational processes and create innovative educational platforms. She can be contacted at email: gulken69@mail.ru.