IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 January 2024, accepted 29 January 2024, date of publication 1 February 2024, date of current version 9 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3361046

RESEARCH ARTICLE

Analysis of Formal Concepts for Verification of Pests and Diseases of Crops Using Machine Learning Methods

JAMALBEK TUSSUPOV[1], MOLDIR YESSENOVA[1], GULZIRA ABDIKERIMOVA[1], AIDYN AIMBETOV[2], KAZBEK BAKTYBEKOV[3], GULDEN MURZABEKOVA[4], AND ULZADA AITIMOVA[4]

[1]L. N. Gumilyov Eurasian National University, 010000 Astana, Kazakhstan
[2]JSC National Company Kazakhstan Garysh Sapary, 010000 Astana, Kazakhstan
[3]Kazakh-French Space Enterprise Galam LLP, 010000 Astana, Kazakhstan
[4]S. Seifullin Kazakh Agro Technical Research University, 010000 Astana, Kazakhstan

Corresponding author: Moldir Yessenova (moldir_11.92@mail.ru)

This work was supported by the Science Committee of the Ministry of Science and Higher Education, Kazakhstan, under Grant AP19677451.

ABSTRACT This article is devoted to a set of important areas of research: the analysis of formal representations and verification of pests and pathogens affecting crops using spectral brightness coefficients (SBR) for the period from 2021 to 2023. The database contains about 10,000 records covering the growing season, types of diseases and pests, as well as their growth phases in a real coordinate system. The work uses machine learning techniques including logistic regression, extreme gradient boosting (XGBoost), and Vanilla convolutional neural network (CNN) to analyze spectral data and classify the presence of pests and diseases in satellite images. The main goal of the work is to optimize and improve the quality of agricultural productivity through early detection and accurate classification of pests and diseases in the agricultural sector. The results of the study can be applied in the development of innovative agricultural systems that will increase yields, reduce the cost of pest and disease control, and optimize production processes. The conclusions of this work can be used both as scientific and practical recommendations for agricultural enterprises and organizations and for the development of new technologies and programs for automating agricultural processes. The use of machine learning techniques and spectral data analysis promises significant breakthroughs in the agricultural sector, helping to improve the efficiency, sustainability, and quality of crop production.

INDEX TERMS Accuracy metrics, classification, clustering, data verification, machine learning, spectral brightness coefficient.

I. INTRODUCTION

Agriculture plays a key role in ensuring food security and sustainable development of society. However, crop pests and diseases have a major impact on crop yields and product quality [1], [2], threatening the efficiency and productivity of the agricultural sector [3]. Pest and disease control is one of the most important tasks of modern agriculture [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

In recent decades, the incidence of pests and diseases has increased significantly due to climate change, global trade flows, migration of pests, and other factors. In [5], an innovative multimodal deep learning framework called ExViT was proposed to solve the problem of land use and land cover classification (LULC) on remote sensing (RS) data. The authors of the paper propose an extension of the traditional Vision Transformer (ViT), which is popular in image processing, to deal more efficiently with multimodal data. Key ideas include processing multimodal image patches using

parallel ViT branches extended with convolution modules to efficiently exploit spatial and modality-specific information. In addition, the use of a cross-modal attention module has been proposed to exchange information between different modalities. The paper presents experimental results on two multimodal remote sensing datasets (Houston 2013 and Berlin), showing that ExViT outperforms competitors based on transformers and convolutional neural networks. The authors also promise to provide source codes and datasets for further research and reproducibility of the results. The authors of this paper [6] propose a method for training a model that can operate on multimodal data (e.g., imagery, radar data, and LiDAR) without explicit labeling. Their approach involves using a common subspace representation and regularization for efficient learning. It is interesting how such methods can improve data processing in remote sensing scenarios.

The presented study proposes a computational model aimed at achieving high accuracy in the task of classifying agricultural lands and their pests, as well as identifying diseases. The developed model can identify diverse crop species, overcoming the limitations of geographical factors. Moreover, the proposed method is based on preliminary training of the model using relevant data. The developed computational model is an integrated approach that combines advanced machine learning methods. The effectiveness of this model is ensured by a training system on diverse data, including diverse geographical distributions of agricultural land and their characteristics. Thus, the model demonstrates high adaptability to various climatic and geographical conditions. The results obtained indicate a significant increase in the accuracy of classification and detection of diseases in agriculture. This computational model provides an important tool for improving crop health monitoring and timely response to potential problems, helping to improve the sustainability and efficiency of agriculture in various regions of the world. Traditional methods of pest and disease control are often ineffective and can lead to negative consequences, such as the development of pesticide resistance in pests or the accumulation of chemical residues in products. During the study, information about the locations in question in the North Kazakhstan and Akmola regions was received from JSC ≪National Company Kazakhstan Garysh Sapary≫. The data covers the period 2021-2023. The data consists of the growing season, types of diseases and pests, and their growth phases in a real coordinate system. The analysis revealed that the database contains 23 types of diseases, pests, and weeds. The database contains about 10,000 data. About 10,000 data points were trained using eXtreme Gradient Boosting (XGBoost), logistic regression, and convolutional neural network machine-learning techniques. To optimize and improve agricultural productivity, developing innovative, sustainable, and effective pest and disease management methods is important [7]. One promising approach is the use of spectral brightness coefficients (SBC) to verify and monitor the condition of crops [8]. Sentinel-2's Vegetation Red Edge channels [9], [10] are an important tool for identifying crop pests and diseases based on **spectral** brightness coefficients (SBC). Sentinel-2 is a space-based Earth observation system mission providing data in a variety of spectral bands (Table 1). Sentinel-2's Vegetation Red Edge channels are in the red edge region of the spectrum, which lies between the Red and Near-Infrared regions. This spectral range is especially informative for analyzing the state of vegetation and identifying changes in crops associated with pests and diseases. Monitoring vegetation health is the ''Vegetation Red Edge'' channels, which allow you to analyze the green chlorophyll characteristics of plants, which helps determine their health and possible stress. If pests or diseases are present, changes in the spectral characteristics of plants can be detected, indicating their presence.

TABLE 1. Sentinel-2 satellite channels used.

Spectral Area	Band-Use	Wavelength (nm)	Description
Band 1 – Coastal aerosol		443.9 - 479.1	This range can be used to detect chlorophyll in plants and assess their health. Changes in the spectral signature of this band may indicate stress, including disease and pests.
Band 2 – Blue		490.6 - 516.5	This range is also used to analyze chlorophyll in plants and can help determine the health of crops
Band 3 – Green		559.8 - 597.6	The green stripe is used to assess the condition and health of plants. Changes in this spectral range may also indicate stress in crops
Band 4 – Red		650.0 - 680.0	The red spectral range is useful for analyzing plant pigmentation and assessing their physiological state
Band 5 – Vegetation red edge		698.5 - 712.5	This range can be used to analyze the condition of plants and identify stress factors
Band 6 – Vegetation red edge		733.5 - 747.5	This range can help identify diseases and stress in plants
Band 7 – Vegetation red edge		773.1 - 793.1	This range is also useful for plant health analysis and disease detection
Band 8 – NIR		793-842	Primary applications include vegetation health analysis and assessment of plant biophysical parameters

Use of Vegetation Red Edge channel spectral data and vegetation indices such as Normalized Difference Vegetation Index (NDVI) [11], [12], [13], Normalized Difference Water Index (NDWI) [14], [15], [16], Soil-Adjusted Vegetation Index (SAVI) [17], [18], [19], Atmospherically Resistant Vegetation Index (ARVI) [20], [21], [22], Enhanced Vegetation Index (EVI) [23], [24], [25], and Green Normalized Difference Vegetation Index (GNDVI) [26], [27], [28], are important tools for monitoring and analyzing the state of vegetation on the Earth's surface. These indices are based on the analysis of spectral data obtained from remote sensing and assess various aspects of vegetation, such as its health, density, and moisture. The NDVI (1) is a numerical metric used in the field of Earth remote sensing to assess and monitor

the health of vegetation on the Earth's surface. NDVI measures the difference in the reflection of light across different spectral ranges, which allows us to determine the presence and density of plant cover

$$NDVI = (NIR - Red)/(NIR + Red) \quad (1)$$

where Near Infrared (NIR) and Red represent reflected values in the near-infrared and red spectral ranges, respectively. NDVI is an index that helps assess the density and health of vegetation. The NDWI is a numerical metric used in the field of Earth remote sensing to detect water surfaces and estimate soil moisture and vegetation on the Earth's surface (2). NDWI can detect the presence of water in remote sensing images and estimate its distribution

$$NDWI = (Green - NIR)/(Green + NIR) \quad (2)$$

where Green represents the reflected value in the green spectral range. NDWI is used to detect water surfaces and estimate soil moisture. The Soil-Adjusted Vegetation Index (SAVI) is an index used in the field of Earth remote sensing to assess and monitor the health of vegetation (3). It is an improved version of the NDVI, which takes into account the influence of soil type on spectral data and allows a more accurate assessment of vegetation health

$$SAVI = ((NIR - Red)/(NIR + Red + L)) * (1 + L) \quad (3)$$

where L is a correction parameter that takes into account the background brightness of the soil. SAVI is designed to account for the influence of soil on spectral data and assess vegetation health. The Atmospherically Resistant Vegetation Index (ARVI) is an index designed to more accurately assess the health of vegetation on the Earth's surface while minimizing the impact of atmospheric conditions and atmospheric distortion on spectral data. ARVI (4) is an improved version of the Normalized Difference Vegetation Index (NDVI) that takes into account the influence of atmospheric conditions on reflected light and helps more accurately assess the health of vegetation

$$ARVI = (NIR - (2 * Red) + Blue)/(NIR + (2 * Red) + Blue) \quad (4)$$

where Blue represents the reflected value in the blue spectral range. ARVI is designed to reduce the influence of atmospheric conditions on spectral data and increase sensitivity to changes in vegetation. The Enhanced Vegetation Index (EVI) is a numerical metric used in the field of Earth remote sensing to assess the health of vegetation on the planet's surface. EVI (5) is an improved version of the NDVI and is designed to account for some of the limitations and distortions of NDVI, such as the effects of atmospheric conditions and surface reflections.

$$EVI = 2.5 * ((NIR - Red)/(NIR + 6 * Red - 7.5 * Blue + 1)) \quad (5)$$

EVI is an improved version of NDVI that also takes into account atmospheric influences and allows for a more accurate assessment of vegetation health. The GNDVI is a numerical metric used in the field of remote sensing to assess the health of vegetation on the Earth's surface. GNDVI (6) measures the difference in the reflection of light in the green spectral range and the near-infrared range, which allows us to determine the presence and density of plant cover.

$$GNDVI = (NIR - Green)/(NIR + Green) \quad (6)$$

GNDVI is similar to NDVI but uses a green value instead of red. This index also evaluates vegetation density. These indices are important tools for monitoring ecosystems, crops, and the environment, and they provide information on the condition and quality of vegetation through the analysis of spectral data obtained from space and other remote platforms. Spectral radiance coefficients reflect electromagnetic radiation measured at different wavelengths and can provide information about the physiological state of plants, the presence of pests and diseases, and stress conditions in crops. Modern technologies for remote sensing of the Earth, such as satellite imagery and unmanned aerial vehicles (drones), make it possible to obtain spectral data with high accuracy and spatial resolution. This work will examine the potential of using spectral radiance coefficients for the verification of crop pests and diseases, as well as their role in optimizing and improving the quality of agricultural productivity. Various methods for analyzing CSCs, such as machine learning and image processing methods, will be considered, as well as examples of the successful implementation of these methods in practice. The results obtained will allow us to conclude the prospects and applicability of spectral brightness coefficients for solving pressing problems in agriculture and ensuring the sustainable development of this important sector of the economy.

II. METHODS

To identify pests and diseases of crops using spectral brightness coefficients (SBC) in the Vegetation Red Edge channels of the Sentinel-2 satellite, various methods of data analysis and processing can be used. Some of these include vegetation indices such as NDVI or NDRE (Normalized Difference Red Edge), calculated from data from the Vegetation Red Edge and Sentinel-2 infrared channels. Recognizing pests and weeds [29] can be challenging and may require the use of machine-learning techniques. However, certain changes in vegetation indices may indicate the presence of pests or weeds. A combination of ranges was used to create a specialized index and the resulting data was processed to identify the unique characteristics of pests and weeds (7).

$$CustomIndex = (RedEdge - Green)/(NIR + Blue + Red) \quad (7)$$

Then, threshold values of this index were used to identify pests and weeds, and these indices were used as features to

train the machine learning model. For pest and weed detection, a machine-learning model was trained using different vegetation indices and spectral bands as features, and pest and weed presence labels as target variables. These indices allow you to assess the health and physiological state of plants. Changes in index values may indicate the presence of stress caused by pests or diseases. Spectral signature analysis is the process of comparing the spectral characteristics of vegetation in infected and healthy crop areas. Changes in spectral characteristics can be detected using spectral curves and help highlight diseased areas. Using machine learning techniques such as classification methods (e.g. Logistic Regression [30], [31], XGBoost [32], [33], Vanilla CNN [34], [35], [36]) can automatically identify diseased or contaminated crop areas based on data from Vegetation Red Edge feeds. This requires training datasets that include information about healthy and contaminated sites. This method uses mathematical algorithms to extract pest or disease characteristics based on changes in vegetation spectra. Spectral data from the Vegetation Red Edge channels can be analyzed using special models to identify signals associated with the presence of pests or diseases. The use of the above methods makes it possible to identify and monitor pests and diseases of crops based on spectral data from the ''Vegetation Red Edge'' channels of the Sentinel-2 satellite [37]. These methods provide more accurate and automated detection of problems in the fields, which helps optimize and improve the quality of agricultural productivity.

Within the framework of this study, the classification of agricultural land plots into normal and anomalous was carried out. The main emphasis was placed on the analysis of various spectral and geographical characteristics, including latitude and longitude coordinates, observation date, and spectral indices B01-B08, NDVI, SAVI, NDWI, ARVI, EVI, and GNDVI. These parameters, collected using satellite or aerial imaging systems, provide valuable information for monitoring land conditions. Three different machine learning models were applied to solve the classification problem: XGBoost, Vanilla CNN, and Logistic Regression. Validation and evaluation of the effectiveness of each model were carried out using metrics such as Accuracy, ROC AUC, Log Loss, F1 Score, Precision, and Recall.

III. RESULTS AND DISCUSSIONS

During the study period, several key data were identified covering various aspects of agricultural management and crop monitoring. Among the identified factors are creeping bitterweed, gypsy moth, dodder, Asian longhorned beetle, as well as important diseases and pests such as Pepino mosaic virus, tomato spotted wilt virus, golden potato nematode, potato canker, grasshopper, wheat thrips, gray armyworm, Asian locust, brown rust, Hessian fly, Italian locust, Colorado potato beetle, meadow moth, mouse rodents, septoria and ground squirrels. Each of the identified factors is characterized by a feature vector that includes 14 key features. These parameters include B01-B08 (visible and infrared reflectance), NDVI (normalized differential vegetation index), SAVI (advanced soil vegetation index), NDWI (normalized water index), ARVI (regional vegetation index), EVI (enhanced vegetation index) and GNDVI (normalized vegetation discrimination index). These trait vectors provide a comprehensive approach to characterizing and monitoring crop health, providing deeper analysis and more accurate predictions regarding the impact of factors influencing crop success and health.

The Pest Verification Model is a comprehensive and structured approach to identifying and confirming the presence of pests in crops. This method involves several sequential steps and applies various techniques to ensure high accuracy and efficiency of the process. The first step uses data from Sentinel-2 satellites and other sources to analyze the spectral characteristics of vegetation. This analysis aims to identify anomalies associated with the presence of harmful organisms. Integration of pest and pathogen information into databases plays a key role in collating and analyzing the resulting data. The application of machine learning algorithms such as classification, clustering, and neural networks acts as an automated tool for analyzing spectral data. These algorithms identify characteristic features that indicate the possible presence of pests. Image processing techniques are widely used to highlight and analyze areas susceptible to potential pest impacts. The process of data verification and classification is complemented by the use of expert systems. These systems play an important role in validating the results obtained from machine learning algorithms, taking into account additional expert information and context. Ultimately, a system is developed for automated decision-making based on the results of data analysis. This step allows us to create an effective mechanism for responding to identified threats, based on objective data and expert examination. Figure 1 shows the pest verification model provides an integrated and science-based approach to protecting farmland from potential threats.

Spectral radiance coefficients are numerical characteristics that describe how objects or surfaces interact with electromagnetic radiation in different parts of the spectrum. To distinguish between weeds and crop species, spectral characteristics can be used to analyze reflected or emitted light. In this study, classification based on spectral characteristics was used to identify weeds or crop species (Figure 2). This allows land cover analysis to be carried out using machine learning or image processing techniques. However, the accuracy of such analyses depends on many factors, including crop diversity, development stages, climatic conditions, and the quality of remote sensing imagery. These factors interact to create a complex context that must be taken into account when interpreting research findings.

To visualize the performance of the classification model, a Confusion Matrix was calculated for all selected three machine learning models (Figure 3). 31 data samples were used to validate the training. This matrix allows you to evaluate how each model copes with the classification of different classes and identify possible errors. Confusion matrix

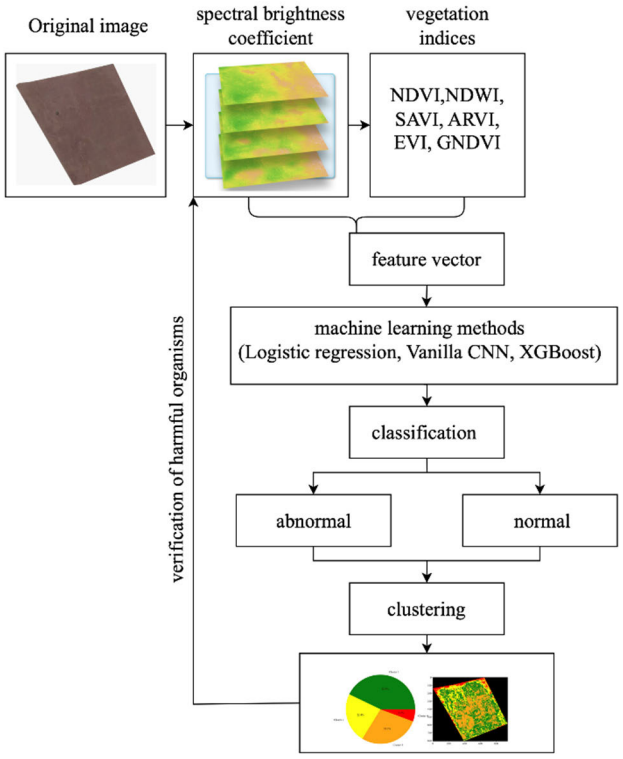

FIGURE 1. Malicious organism verification model.

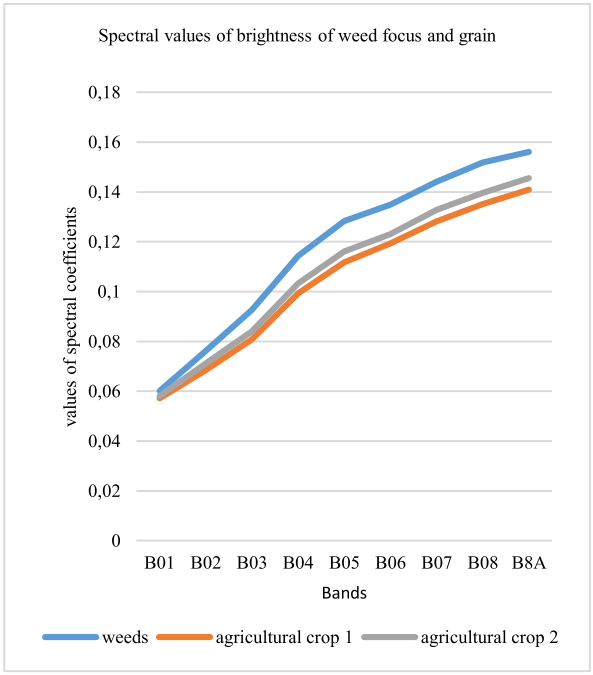

FIGURE 2. Verification of harmful organisms using spectral brightness coefficients.

analysis helps to better understand the strengths and weaknesses of models in the context of the selected problem. When calculating the error matrices in the context of the machine learning method, logistic regression showed four

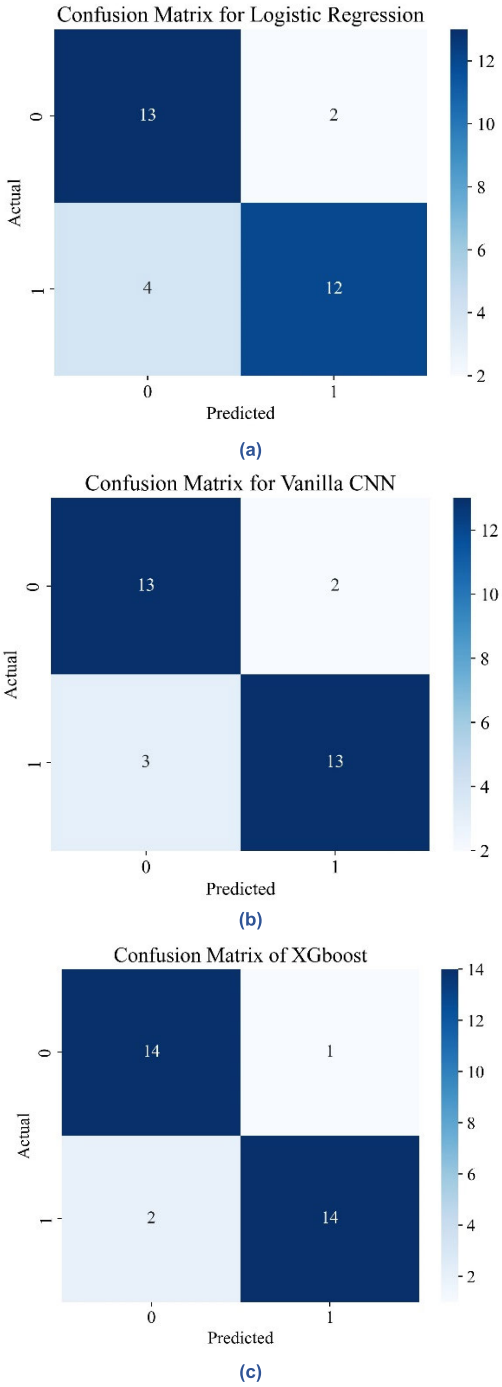

FIGURE 3. Confusion matrix for machine learning methods.

cases of misclassification (Figure 3(a)). At the same time, the XGBoost method revealed two errors of the second type (Figure 3(c)). These results indicate differences in the performance of the two models, where logistic regression made more Type I errors while XGBoost made fewer Type II errors. Analyzing such discrepancies in the error matrix provides valuable information for further tuning and optimization of models taking into account the specifics of the classification problem. As shown in Figure 3(b), the number of examples that were correctly classified as positive using the Vanilla

CNN method was 13, while the XGBoost method detected 14 such examples (Figure 3(c)). These results reflect the effectiveness of both models in identifying positive classes, where XGBoost showed a slight increase in accuracy compared to Vanilla CNN. Analyzing such differences provides information about the comparative performance of methods and can serve as a basis for further tuning of models.

In this work, the XGBoost method proved to be the more efficient model among those considered, but it is necessary to carefully monitor its performance during testing. Vanilla CNN, although providing good results, requires optimization for better generalization. Logistic regression, in turn, was less effective. Based on the provided data and calculations, the Matthews Correlation Coefficient (MCC) values reveal the strength of association between predictions and actual values for the three models. Notably, XGBoost exhibits the highest MCC, indicating its superior predictive capability ($MCC\_LR \approx 0.445$, $MCC\_CNN \approx 0.612$, $MCC\_XGBoost \approx 0.722$). The Fowlkes-Mallows Index, reflecting similarity, underscores the models' performance in achieving agreement. Both XGBoost and Logistic Regression with XGBoost showcase a proximity to unity in their similarity indices, denoting a high degree of concordance ($FM\_LR\_CNN \approx 0.827$, $FM\_LR\_XGBoost \approx 0.860$, $FM\_CNN\_XGBoost \approx 0.852$). Considering Cohen's Kappa Coefficient, a metric accounting for agreement while adjusting for chance, XGBoost consistently outperforms the other models, reinforcing its effectiveness in predictions ($Kappa\_LR \approx 0.391$, $Kappa\_CNN \approx 0.562$, $Kappa\_XGBoost \approx 0.660$).

In summary, the elevated MCC values signify robust predictive relationships, with XGBoost demonstrating the most substantial association. Similarly, the high Fowlkes-Mallows Index values underscore strong agreement, particularly evident in the comparisons involving XGBoost. Cohen's Kappa Coefficient further substantiates the superior performance of XGBoost in achieving both agreement and adjustment for chance. Collectively, these findings suggest XGBoost's effectiveness and precision in prediction, distinguishing it from Logistic Regression and Vanilla CNN in this analysis. Figure 4 shows the comparative result of all three models, which provides important indicators for choosing a model and possible directions for its improvement, taking into account a specific classification task.

Using the above 14 features, a database containing 15,000 entries for each machine-learning method was constructed. The data presented in the database has a multidimensional structure, reflecting various aspects associated with the phenomena of agricultural land and their interaction with pests, diseases, and other factors. The model training process was carried out based on the provided data using various machine learning methods. Each method was trained on a sample of 15,000 records, allowing the model to learn a variety of patterns and dependencies in the data. As a result of training, the model acquired the ability to classify, making it possible to determine the classification accuracy for each of the considered machine learning methods.

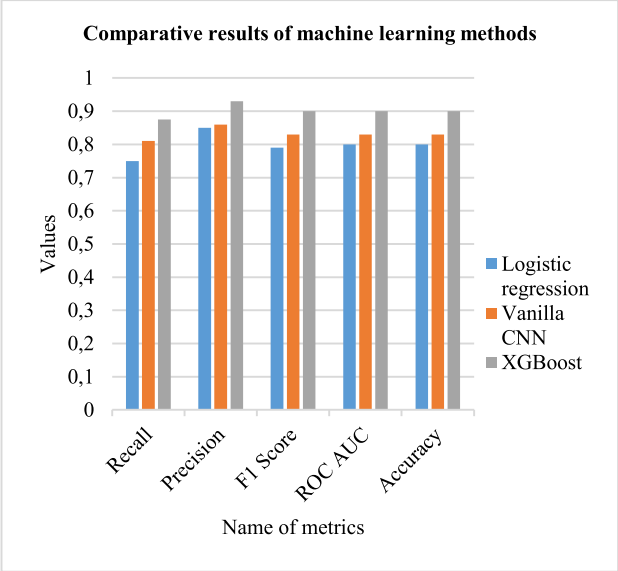

FIGURE 4. Comparative results of machine learning methods.

The XGBoost model demonstrated exceptional performance, achieving accuracy and ROC AUC scores of 90.32% and 0.9, respectively. The Log Loss value of 3.48, and the maximum values of F1 Score, Precision, and Recall equal to 0.9, indicate the high reliability of the model. This is especially significant given the complexity of the task and the variety of characteristics based on which the classification was based. When testing training, the XGBoost method showed high accuracy - 97.98%. Such efficiency can be attributed to the algorithmic features of XGBoost, which make this model extremely adaptive to various data types and structural anomalies. When tested, the model also demonstrated 90% accuracy, confirming its applicability and reliability.

The use of convolutional neural networks (Vanilla CNN) also showed good results, although slightly inferior to XGBoost. With a training accuracy of 83.87%, ROC AUC of 0.83, and F1 Score of 0.838, this model provides reliable classification. However, it is worth noting that the accuracy during testing decreased to 71.43%. This may indicate overtraining of the model or insufficient generalization, which requires additional analysis and possible correction of the architecture or training parameters. Although CNNs are commonly used for image analysis, their successful application in this study highlights their versatility and ability to identify complex relationships in data.

The logistic regression model performed less impressively than other models, with a training accuracy of 80% and an ROC AUC of 0.808. These figures, together with the F1 Score of 0.799 and Log Loss of 6.97, indicate the relatively poor performance of this model for this task. At the same time, the accuracy during testing was 71.43%, which is close to the training indicators and may indicate sufficient stability of the model. However, given the lack of performance, optimization

of the model or other approaches should be considered to improve classification performance. According to the growing season data of 06/01/2023, the anomalous value was determined and clarified by specialists at the coordinate point of latitude - 69.888 and longitude - 52.336. The model was tested on this piece of land. The result was a weed outbreak of 5.7 percent, as shown in Figure 5.

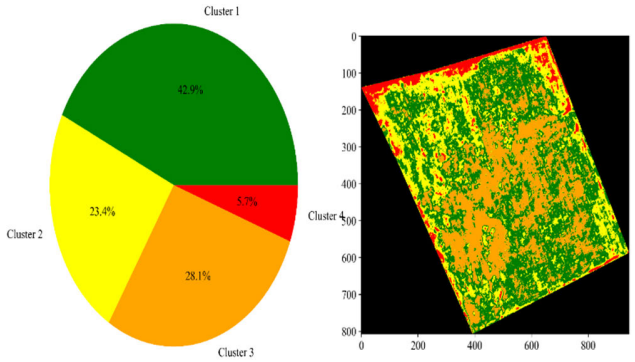

FIGURE 5. Clustering result using the XGBoost model.

By comparing the performance of three different models - XGBoost, Vanilla CNN, and Logistic Regression - in the context of the agricultural plot classification task, we can come to a convincing conclusion about the outstanding performance of the XGBoost model.

In this study, an in-depth analysis and experimental verification of the time complexity of the XGBoost algorithm used for classification problems was carried out. The XGBoost algorithm was chosen for its efficiency and wide application in modern machine learning. To estimate time complexity, data consisting of 1707 observations with 14 features, including numerical variables and time stamps, was used. As a first step, the data was preprocessed, including converting timestamps into a numeric format relative to a given start date. To select the optimal hyperparameters of the model, the GridSearchCV method was used, which made it possible to fine-tune the model to the specifics of the data. The main criterion for assessing the effectiveness of the model was the accuracy metric, and additional metrics were also considered, such as F1-score, precision, and recall, to more fully understand the performance of the model. Experimental testing involved measuring the time required to train the model and subsequently make predictions. This made it possible to estimate the time complexity of the algorithm. The results showed that the training time increases linearly with the size of the input data, which is consistent with the theoretical assumptions of the linear time complexity of the XGBoost algorithm. Thus, it was experimentally confirmed that the algorithm has linear time complexity in the context of the data presented. Finally, the research methodology, detailed data analysis, model training, and evaluation process, as well as the timing characteristics of the algorithm are documented. The study's findings highlight the effectiveness and scalability of XGBoost in machine learning tasks, especially when working with large data sets, confirming its suitability for a wide range of data science applications.

This model demonstrated the best results in all metrics considered and confirmed its effectiveness during testing. XGBoost has high classification accuracy, reliability, and stability in predictions. Its ability to work with various types of data, including numerical and categorical attributes, as well as time series processing, makes it a powerful tool for analyzing and classifying agricultural areas.

On the other hand, Vanilla CNN and Logistic Regression also performed well, but they were less stable compared to XGBoost. These models may require additional optimization and parameter tuning to achieve high performance, and their application may be more complex in the context of agronomic data. The obtained results of the study have important implications for application in real conditions. They can be used to monitor and manage agricultural plots, which will help agricultural enterprises and agronomists make informed decisions about crop care and pest control. In addition, the methodology applied in this study can be adapted and applied to other classification problems in the fields of agronomy and ecology, thereby expanding its scope of applicability and impact. This is an important step towards optimizing and making agriculture more sustainable in the modern world.

This approach allows us to systematically evaluate the performance of different machine learning methods in the context of agricultural problems. The obtained classification accuracies are a key criterion for comparative analysis of the effectiveness of methods, contributing to a better understanding and selection of optimal approaches to solving problems in the field of agricultural technologies.

IV. CONCLUSION

In conclusion of this article, it can be noted that the study of verification of harmful organisms and pathogens affecting crops using spectral-space data and machine learning methods, including logistic regression, Vanilla CNN, and XGBoost, is a relevant and promising area of research. These methods and tools have a significant impact on optimizing and increasing agricultural productivity in modern conditions.

Logistic regression, Vanilla CNN, and XGBoost demonstrate their effectiveness in various aspects of the problem of verifying harmful organisms and pathogens. Logistic regression provides a simple and interpretable method for classifying and assessing the risks associated with pest exposure to crops. On the other hand, Vanilla CNN can efficiently work with image and spectral data, making it a powerful tool for automatically identifying and classifying pathogens and harmful organisms. This method helps identify even complex samples and allows analysis at a higher level of detail. XGBoost, on the other hand, stands out for its ability to work efficiently with numerical data and time series. This boosting algorithm helps improve the accuracy and reliability

of forecasts and can be used to more accurately estimate the proportion of harmful organisms and pathogens.

In conclusion, the combined use of spectral-space data and machine learning techniques in agriculture has enormous potential to improve the sustainability of agricultural systems, optimize resources, and increase productivity. These innovative approaches lead to more precise and efficient farmland management, which is critical in the face of growing populations and changing climate conditions.

ACKNOWLEDGMENT

The authors would like to thank the Joint Stock Company ''National Company'' ''Kazakhstan Gharysh Sapary'' for their valuable contribution to this research. The provision of data from this organization was critical to the successful conduct of research aimed at identifying and verifying harmful organisms and pathogens and optimizing and increasing productivity in crops.

REFERENCES

[1] M. Yessenova, G. Abdikerimova, A. Adilova, A. Yerzhanova, N. Kakabayev, T. Ayazbaev, Z. Sattybaeva, and T. Ospanova, ''Identification of factors that negatively affect the growth of agricultural crops by methods of orthogonal transformations,'' *Eastern-Eur. J. Enterprise Technol.*, vol. 117, no. 2, pp. 39–47, Jun. 2022.

[2] M. Yessenova, G. Abdikerimova, N. Baitemirova, G. Mukhamedrakhimova, K. Mukhamedrakhimov, Z. Sattybaeva, I. Salgozha, and A. Yerzhanova, ''The applicability of informative textural features for the detection of factors negatively influencing the growth of wheat on aerial images,'' *Eastern-Eur. J. Enterprise Technol.*, vol. 118, no. 2, pp. 51–58, Aug. 2022.

[3] M. Yessenova, G. Abdikerimova, T. Ayazbaev, G. Murzabekova, A. Ismailova, Z. Beldeubayeva, A. Ainagulova, and A. Mukhanova, ''The effectiveness of methods and algorithms for detecting and isolating factors that negatively affect the growth of crops,'' *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 13, no. 2, pp. 1669–1679, Apr. 2023.

[4] M. Yessenova, G. Abdikerimova, Z. B. Sadirmekova, N. Glazyrina, S. Adikanova, A. Tanirbergenov, M. Karipola, and G. Mukhamedrakhimova, ''Features of growth of agricultural crops and factors negatively affecting their growth,'' *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 30, no. 1, pp. 625–632, Apr. 2023, doi: 10.11591/ijeecs.v30.i1.pp625-632.

[5] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, ''Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514415.

[6] J. Yao, D. Hong, H. Wang, H. Liu, and J. Chanussot, ''UCSL: Toward unsupervised common subspace learning for cross-modal image classification,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514212.

[7] G. Abdikerimova, M. Yessenova, A. Yerzhanova, Z. Manbetova, G. Murzabekova, D. Kaibassova, R. Bekbayeva, and M. Aldashova, ''Applying textural Law's masks to images using machine learning,'' *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 13, no. 5, p. 5569, Oct. 2023.

[8] M. Yessenova, G. Abdikerimova, G. Murzabekova, K. Nurbol, N. Glazyrina, S. Adikanova, N. Uzakkyzy, Z. B. Sadirmekova, and R. Niyazova, ''Application of informative textural Law's masks methods for processing space images,'' *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 13, no. 4, p. 4557, Aug. 2023.

[9] J. Li, Z. Wu, Z. Hu, Z. Li, Y. Wang, and M. Molinier, ''Deep learning based thin cloud removal fusing vegetation red edge and short wave infrared spectral information for Sentinel-2A imagery,'' *Remote Sens.*, vol. 13, no. 1, p. 157, Jan. 2021.

[10] S. Lin, J. Li, Q. Liu, L. Li, J. Zhao, and W. Yu, ''Evaluating the effectiveness of using vegetation indices based on red-edge reflectance from Sentinel-2 to estimate gross primary productivity,'' *Remote Sens.*, vol. 11, no. 11, p. 1303, May 2019.

[11] P. V. Lykhovyd, R. A. Vozhehova, S. O. Lavrenko, and N. M. Lavrenko, ''The study on the relationship between normalized difference vegetation index and fractional green canopy cover in five selected crops,'' *Sci. World J.*, vol. 2022, pp. 1–6, Mar. 2022.

[12] L. M. Morales-Gallegos, T. Martínez-Trinidad, P. Hernández-de la Rosa, A. Gómez-Guerrero, D. Alvarado-Rosales, and L. de Lourdes Saavedra-Romero, ''Tree health condition in urban green areas through crown indicators and vegetation indices,'' Tech. Rep., 2023.

[13] J. Janoušek, P. Marcoń, P. Dohnal, V. Jambor, H. Synková, and P. Raichl, ''Predicting the optimum corn harvest time via the quantity of dry matter determined with vegetation indices obtained from multispectral field imaging,'' *Remote Sens.*, vol. 15, no. 12, p. 3152, Jun. 2023.

[14] J. Li, Y. Meng, Y. Li, Q. Cui, X. Yang, C. Tao, Z. Wang, L. Li, and W. Zhang, ''Accurate water extraction using remote sensing imagery based on normalized difference water index and unsupervised deep learning,'' *J. Hydrol.*, vol. 612, Sep. 2022, Art. no. 128202.

[15] E. Sovisoth, V. S. Kuntal, P. Misra, W. Takeuchi, and K. Nagai, ''Estimation of year of construction of bridges in Cambodia by analyzing the Landsat normalized difference water index,'' *Infrastructures*, vol. 8, no. 4, p. 77, Apr. 2023.

[16] V. N. Husna and N. I. Fawzi, ''Aplikasi algoritma normalized difference water index (NDWI), normalized difference vegetation index (NDVI) dan bare soil index (BSI) dalam penilaian kerapatan vegetasi dan produktivitas di Pulau burung,'' *Geo-Image J.*, vol. 11, no. 2, pp. 198–204, 2022.

[17] T. Konno and K. Homma, ''Prediction of areal soybean lodging using a main stem elongation model and a soil-adjusted vegetation index that accounts for the ratio of vegetation cover,'' *Remote Sens.*, vol. 15, no. 13, p. 3446, Jul. 2023.

[18] H. Kareem, M. Attaee, and Z. Omran, ''Evaluation the soil-adjusted vegetation indices SAVI and MSAVI for Bristol city, united kingdom using Landsat 8-OLI through geospatial technology,'' *Ecological Eng. Environ. Technol.*, vol. 24, no. 7, pp. 89–97, 2023.

[19] M. A. G. Aranda, A. M. Talamantes, J. Sánchez-Salas, and D. U. G. Uribe, ''Soil adjusted vegetation index to estimate management effect at North Coahuila,'' *Ecosistemas y Recursos Agropecuarios*, vol. 9, no. 3, 2022.

[20] P. Lemenkova and O. Debeir, ''Computing vegetation indices from the satellite images using GRASS GIS scripts for monitoring mangrove forests in the coastal landscapes of Niger delta, Nigeria,'' *J. Mar. Sci. Eng.*, vol. 11, no. 4, p. 871, Apr. 2023.

[21] S. Asgari and M. Hasanlou, ''A comparative study of machine learning classifiers for crop type mapping using vegetation indices,'' *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 10, pp. 79–85, Jan. 2023.

[22] B. Bayaraa, A. Hirano, M. Purevtseren, B. Vandansambuu, B. Damdin, and E. Natsagdorj, ''Applicability of different vegetation indices for pasture biomass estimation in the north-central region of Mongolia,'' *Geocarto Int.*, vol. 37, no. 25, pp. 7415–7430, Dec. 2022.

[23] H. Wang, L. He, J. Yin, Z. Yu, S. Liu, and D. Yan, ''Effects of effective precipitation and accumulated temperature on the terrestrial EVI (enhanced vegetation index) in the yellow river basin, China,'' *Atmosphere*, vol. 13, no. 10, p. 1555, Sep. 2022.

[24] C. Wang, Y. Wu, Q. Hu, J. Hu, Y. Chen, S. Lin, and Q. Xie, ''Comparison of vegetation phenology derived from solar-induced chlorophyll fluorescence and enhanced vegetation index, and their relationship with climatic limitations,'' *Remote Sens.*, vol. 14, no. 13, p. 3018, Jun. 2022.

[25] N. Mehra and J. B. Swain, ''Use of enhanced vegetation index (EVI) as a land cover classification tool and its suitability in urban development studies,'' in *Proc. AIP Conf.*, vol. 2852, 2023, Art. no. 130004.

[26] B. E. Cahyono, A. T. Nugroho, and N. D. Wulandari, ''Analisis usia tebu terhadap pola nilai gndvi (green normalized difference vegetation index) berdasarkan data citra Landsat-8,'' *J. Teknotan*, vol. 16, no. 3, pp. 139–146, Dec. 2022.

[27] M. H. Khudhur, I. A. Alwan, and N. A. Aziz, ''Comparison of the accuracies of different spectral indices for mapping the vegetation covers in Al-Hawija district, Iraq,'' in *Proc. AIP Conf.*, vol. 2775, 2023, Art. no. 050005.

[28] P. Pipatsitee, R. Tisarum, K. Taota, T. Samphumphuang, A. Eiumnoh, H. P. Singh, and S. Cha-um, ''Effectiveness of vegetation indices and UAV-multispectral imageries in assessing the response of hybrid maize (*Zea mays* L.) to water deficit stress under field environment,'' *Environ. Monitor. Assessment*, vol. 195, no. 1, Jan. 2023.

[29] H. Obasekore, M. Fanni, S. M. Ahmed, V. Parque, and B.-Y. Kang, "Agricultural robot-centered recognition of early-developmental pest stage based on deep learning: A case study on fall armyworm (*Spodoptera frugiperda*)," *Sensors*, vol. 23, no. 6, p. 3147, Mar. 2023.
[30] K. Arumugam, Y. Swathi, D. T. Sanchez, M. Mustafa, C. Phoemchalard, K. Phasinam, and E. Okoronkwo, "Towards applicability of machine learning techniques in agriculture and energy sector," *Mater. Today, Proc.*, vol. 51, pp. 2260–2263, Jan. 2022.
[31] H. Pallathadka, M. Jawarneh, F. Sammy, V. Garchar, D. T. Sanchez, and M. Naved, "A review of using artificial intelligence and machine learning in food and agriculture industry," in *Proc. 2nd Int. Conf. Advance Comput. Innov. Technol. Eng. (ICACITE)*, Apr. 2022, pp. 2215–2218.
[32] H. Yan, Z. He, C. Gao, M. Xie, H. Sheng, and H. Chen, "Investment estimation of prefabricated concrete buildings based on XGBoost machine learning algorithm," *Adv. Eng. Informat.*, vol. 54, Oct. 2022, Art. no. 101789.
[33] P. A. C. Rocha and V. O. Santos, "Global horizontal and direct normal solar irradiance modeling by the machine learning methods XGBoost and deep neural networks with CNN-LSTM layers: A case study using the GOES-16 satellite imagery," *Int. J. Energy Environ. Eng.*, vol. 13, no. 4, pp. 1271–1286, Dec. 2022.
[34] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Comput. Electron. Agricult.*, vol. 177, Oct. 2020, Art. no. 105709.
[35] S. Khan and M. Narvekar, "Disorder detection of tomato plant(solanum lycopersicum) using IoT and machine learning," *J. Phys., Conf.*, vol. 1432, no. 1, Jan. 2020, Art. no. 012086.
[36] A. Ramanath, S. Muthusrinivasan, Y. Xie, S. Shekhar, and B. Ramachandra, "NDVI versus CNN features in deep learning for land cover clasification of aerial images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 6483–6486.
[37] A. Ivanda, L. Serić, M. Bugarić, and M. Braović, "Mapping chlorophyll-a concentrations in the kaštela bay and Brač channel using ridge regression and Sentinel-2 satellite images," *Electronics*, vol. 10, no. 23, p. 3004, Dec. 2021.

JAMALBEK TUSSUPOV received the Graduate degree in mathematics from the Faculty of Mathematics, Karaganda State University, in 1979, and the Ph.D. degree in mathematics from Novosibirsk State University, in 1990. He continued his education in graduate school with KarSU and was sent to the Institute of Mathematics named after. S. L. Soboleva, Siberian Branch of the USSR Academy of Sciences, Novosibirsk, from 1983 to 1986. He defended his Ph.D. dissertation from Al-Farabi Kazakh National University, in 2007, on the topic "Problems of definability and algorithmic complexity of relations over algebraic structures." He completed a Scientific Internship from the University of Notre Damme, IN, USA, in 2011, being a scholarship recipient of the Bolashak Presidential Program for this period. In 2011, he was awarded the title of "Professor" by the Ministry of Education and Science, Kazakhstan.

MOLDIR YESSENOVA received the bachelor's degree in information systems and the master's degree in information technology from L. N. Gumilyov Eurasian National University, Astana, Kazakhstan, in 2014 and 2017, respectively, where she is currently pursuing the Ph.D. degree with the Department of Information Systems. Her research interests include image processing, computer vision, satellite imagery, artificial intelligence, and machine learning.

GULZIRA ABDIKERIMOVA received the Ph.D. degree in information systems from L. N. Gumilyov Eurasian National University, Kazakhstan, in 2020. Currently, she is an Associate Professor with the Department of Information Systems, L. N. Gumilyov Eurasian National University. Her research interests include image processing, computer vision, satellite imagery, artificial intelligence, and machine learning.

AIDYN AIMBETOV is currently a Cosmonaut-Researcher, has experience flying on the International Space Station. He is also the Chairman of the Management Board of JSC NC "Kazakstan Garysh Sapary," which specializes in the field of remote sensing of the earth. On his initiative with Eurasian National University named after L.N. Gumilyov for his Ph.D. studies in space engineering have been opened, where he lectures with the Department of Space Engineering and Technology on Astronautics. His research interests include astronautics and remote sensing of the earth. He has 13 publications (four in the Scopus database) and three patents for inventions over the past five years.

KAZBEK BAKTYBEKOV has teaching experience at leading universities in Kazakhstan and (National Research Nuclear University "MEPhI," Moscow). Currently, he is a leading Design Engineer with joint Kazakh-French Space Enterprise Ghalam LLP. He is also the Head of Grant Research Projects. Over the past five years, he has published 17 scientific articles, including nine in journals included in the Scopus database. His research interests include nonlinear, fractal physics, remote sensing of the earth, and satellite engineering, under the guidance of Prof. K. Baktybekov.

GULDEN MURZABEKOVA received the Graduate degree in applied mathematics from St. Petersburg State University, in 1994. In December 1997, she defended her Ph.D. thesis in the specialty 01.01.09—discrete mathematics and mathematical cybernetics. Her research interests include non-smooth analysis, non-differentiable optimization, and computer modeling and machine learning.

ULZADA AITIMOVA received the Candidate degree in physical and mathematical sciences from KazATK named after M. Tynyshpaev, in December 2000. After receiving the degree, she focused on developing and analyzing computer models, designing information systems, digitalizing IT infrastructure, and recognizing images using artificial intelligence. Over the past five years, ten scientific articles have been published, including three in journals included in the Scopus database. Her research interests include computer modeling, image analysis, and machine learning in relation to digitalization of agricultural systems.