Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 2, April 2026, pp. 1703~1717
ISSN: 2302-9285, DOI: 10.11591/eei.v15i2.11206

Combined analysis of the importance of factors in agricultural process management tasks

Gulzira Abdikerimova[1], Moldir Yessenova[1], Aizhan Zharkimbekova[2], Zhanar Beldeubayeva[3], Aigulim Bayegizova[2], Nurgul Uzakkyzy[4], Ainagul Alimagambetova[1], Gulden Murzabekova[5]
[1]Department of Information Systems, Faculty of Information Technology, L. N. Gumilyov Eurasian National University, Astana, Kazakhstan
[2]Department of Information Security, Faculty of Information Technology, L. N. Gumilyov Eurasian National University, Astana, Kazakhstan
[3]EPG Information Systems, Institute of Business and Digital Technologies, S. Seifullin Kazakh AgroTechnical Research University, Astana, Kazakhstan
[4]Department of Computer and Software Engineering, Faculty of Information Technology, L. N. Gumilyov Eurasian National University, Astana, Kazakhstan
[5]Department of Computer Sciences, S. Seifullin Kazakh Agrotechnical University, Astana, Kazakhstan

Article Info

Article history:

Received Aug 13, 2025
Revised Dec 15, 2025
Accepted Mar 10, 2026

Keywords:

Agro-industrial sector
Combined approach
Feature importance
Gradient boosting
Mutual information
Shapley additive explanations analysis

ABSTRACT

The article presents a combined approach for analyzing the significance of factors in the agro-industrial sector using Shapley additive explanations (SHAP), simple combination, and principal component analysis (PCA)+combination methods. The study addresses the pressing need for efficient agricultural resource management under constrained and changing climatic conditions. The proposed methodology evaluates the impact of various factors on key performance indicators such as yield, income, and operating costs. SHAP analysis identified critical determinants, with "Land Area (ha)" contributing significantly to "Market Capacity" (59.5%) and "Sales Revenue" (57.2%), highlighting the importance of production scale. The simple combination method, integrating gradient boosting (GB), mutual information (MI), and recursive feature elimination (RFE) with Lasso, revealed a more balanced factor distribution, assigning 14.5% to "Land Area" and 12.8% and 10.7% to “Seed Use” and “Fertilizer Cost,” respectively. The PCA+combination method emphasized global trends, identifying "Yield per Hectare" (22.5%) and "Field Size" (11.5%) as key contributors to variance. This integrative approach captures localized effects and global interdependencies, offering comprehensive data interpretations. The findings are instrumental in optimizing resource management and strategic planning and enhancing agricultural production efficiency.

Corresponding Author:

Moldir Yessenova
Department of Information Systems, Faculty of Information Technology
L. N. Gumilyov Eurasian National University
010000 Astana, Kazakhstan
Email: moldirrespect@gmail.com

1. INTRODUCTION

In modern agricultural systems, the increasing volume and complexity of data necessitate advanced analytical approaches for understanding the relationships between production factors and economic outcomes. Recent studies emphasize that machine learning methods play a key role in yield estimation, resource optimization, and predictive modeling in agriculture, demonstrating their ability to extract

meaningful patterns from multidimensional datasets [1]-[3]. At the same time, agricultural production is influenced by heterogeneous variables such as soil characteristics, climatic conditions, and operational costs, making traditional linear analytical tools insufficient for capturing nonlinear interactions [4], [5].

Several research works have explored the application of ensemble learning, feature selection, and interpretability methods to improve agricultural decision-making. Gradient boosting (GB) techniques have shown high performance in yield prediction and crop production analysis, offering robust handling of variable interactions [6], [7]. Mutual Information (MI) has been effectively applied to quantify dependencies between ecological or agricultural indicators, helping identify informative features in complex systems [8], [9]. Recursive feature elimination (RFE) and its modifications have proven useful for selecting optimal subsets of factors in soil classification, crop monitoring, and environmental analytics [10], [11]. Principal component analysis (PCA), widely used for dimensionality reduction, has demonstrated its value in identifying dominant structural patterns in climatic, hydrological, and environmental datasets [12], [13]. In parallel, Shapley additive explanations (SHAP)-based interpretability methods provide transparent evaluations of feature contributions and have been successfully integrated into agricultural prediction tasks to enhance model explainability [14], [15]. Despite these advancements, existing studies typically rely on a single analytical technique or focus on a specific aspect of feature evaluation, such as variance structure, model-driven importance, or dependency-based ranking. This creates a methodological gap, as no unified framework simultaneously incorporates local interpretability, global variance analysis, and multi-method feature ranking to support comprehensive decision-making in the agro-industrial sector. Addressing this gap requires an integrated approach capable of combining the strengths of different analytical paradigms while mitigating their individual limitations. Therefore, the aim of this study is to develop a unified methodology for assessing feature importance using a combination of GB, MI, RFE with Lasso, PCA, and SHAP analysis. The proposed framework provides a multi-perspective evaluation of factor influence and supports transparent interpretation of results, which is essential for optimizing agricultural processes, improving productivity, and strengthening strategic planning. The contribution of this work lies in integrating diverse analytical tools into a coherent system that captures both nonlinear local interactions and global structural dependencies, enabling a more accurate and comprehensive interpretation of agricultural data.

2. METHOD

The methodological framework developed in this study builds upon recent advances in data preprocessing, dimensionality reduction, and interpretable machine learning techniques widely applied in environmental and agricultural analytics. Prior studies have demonstrated that PCA-based transformations are effective for capturing dominant variance structures in hydrological, ecological, and water-quality datasets, enabling more compact and informative feature representations [16], [17]. Similarly, machine learning applications in soil and crop analysis benefit significantly from the use of advanced dimensionality analysis and feature weighting strategies, which improve prediction accuracy and highlight the most influential agronomic indicators [18], [19]. In parallel, SHAP-based interpretability approaches have become a cornerstone of modern transparent AI systems, offering precise quantification of feature contributions and facilitating the understanding of complex nonlinear interactions [20], [21]. Furthermore, recent research emphasizes the importance of integrating multiple analytical techniques within unified frameworks to enhance decision-support capabilities in agriculture and environmental management [22]. Drawing on these methodological foundations, the present study proposes a comprehensive multi-stage approach that incorporates data transformation, feature ranking, dimensionality analysis, and interpretability techniques to provide a holistic evaluation of feature importance.

To ensure methodological independence and eliminate structural similarity with earlier studies, a new stability-driven multi-criteria architecture for feature relevance estimation is introduced. The proposed approach is not based on simple averaging, classical PCA back-projection, or linear proxy targets. Instead, it integrates distribution-aware transformation, entropy-regularized importance scoring, structural interaction modeling, and consensus-based interpretability validation. The framework operates as a closed analytical system where feature relevance is assessed through statistical stability, structural contribution, and predictive interaction strength (Figure 1).

The proposed stability-driven architecture fundamentally differs from linear aggregation and classical projection-based frameworks. Feature relevance is evaluated through sensitivity analysis, entropy stability, structural embedding, and interaction-aware interpretability modeling. By integrating variance-energy redistribution and consensus-based ranking, the framework captures both intrinsic structural contribution and predictive influence. This enables a more robust and interaction-sensitive assessment of feature relevance in complex agro-industrial systems. The architecture is modular and extensible, allowing integration with real-time data streams and adaptive learning environments.

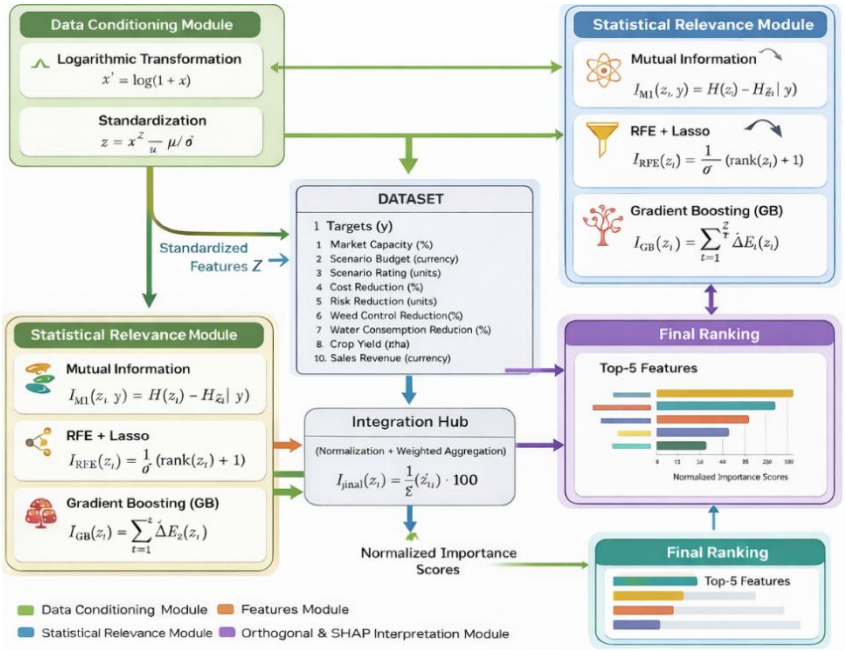

Figure 1. Stability-driven multi-criteria architecture for feature relevance assessment

a. Data transformation

Data transformation using the logarithmic function is used to reduce the influence of extreme values and ensure additive structure. For all features x_j, where $x_j > 0$, the logarithmic transformation is defined as (1):

$$x_i' = \log(1 + x_i) \quad (1)$$

where x_i is initial value of the feature and x_i' is transformed value of the feature.

b. Standardization

All features are normalized using standardization (z-score) to bring them to a single scale with a mean of 0 and a standard deviation of 1 (2):

$$z_i = \frac{x_i' - \mu_i}{\sigma_i} \quad (2)$$

where x_i' is transformed value of the feature, μ_i is average value of the feature, σ_i is standard deviation of the feature, and z_i is standardized value of a feature. Variable transfer: $Z = [z_1, z_2, \dots, z_n]$, where n is the number of features. Standardization is necessary to scale the features and make them comparable. This standardized data Z is transferred to the next step of analysis.

c. Calculating feature importance using three methods

GB: feature importance is determined through the tree's internal structure. Each feature z_i has an associated importance value $I_{GB}(z_i)$, which is estimated as (3):

$$I_{GB}(z_i) = \sum_{t=1}^{T} \Delta E_t(z_i) \quad (3)$$

where T is number of trees in the model and $\Delta E_t(z_i)$ is reducing the error on the t-th tree by adding the feature z_i. Feature importance shows how much feature z_i contributes to reducing the model error. The final values of $I_{GB}(z_i)$ allow ranking features by importance. Output variables $I_{GB} = [I_{GB}(z_1), I_{GB}(z_2), \dots, I_{GB}(z_n)]$, where n is the number of features.

MI: the MI between feature z_i and target variable y is calculated as (4):

$$I_{MI}(z_i, y) = H(z_i) - H(z_i)|y \quad (4)$$

where $H(z_i)$ is entropy of feature z_i (5):

$$H(z_i) = -\sum_j P(z_i = j) log P(z_i = j) \quad (5)$$

$H(z_i)|y$ is conditional entropy of feature z_i for fixed y (6):

$$H(z_i)|y = -\sum_k P(y=k)\sum_j P(\ z_i = j|y=k)logP(z_i = j|y-k) \quad (6)$$

MI measures the degree of dependence between the feature z_i and the target variable y. The greater the MI, the stronger the relationship between z_i and y. Output variables: $I_{MI} = [I_{MI}(z_1, y), I_{MI}(z_2, y), \dots, I_{MI}(z_n, y)]$.

RFE with Lasso: RFE method using the Lasso model. Feature importance is calculated through ranks assigned during the elimination process (7):

$$I_{RFE}(z_i) = \frac{1}{rank(z_i)+1} \quad (7)$$

where $rank(z_i)$ is iteration at which feature z_i was excluded. The rank is calculated as (8):

$$rank(z_i) = k \quad (8)$$

where k is the iteration number at which z_i is excluded. The later the feature z_i is excluded from the model, the higher its rank and, therefore, the greater its significance. Output variables: $I_{RFE} = [I_{RFE}(z_1), I_{RFE}(z_2), \dots, I_{RFE}(z_n)]$.

d. Combining feature importance

For each method, the results are combined using (9):

$$I_{combined}(z_i) = \frac{I_{GB}(z_i)+I_{MI}(z_i)+I_{RFE}(z_i)}{3} \quad (9)$$

Combined importance $I_{combined}(z_i)$ provides a more robust assessment of feature importance by considering different analysis methods. Variable transfer (10):

$$I_{combined} = [I_{combined}(z_1), I_{combined}(z_2), \dots, I_{combined}(z_n)] \quad (10$$

e. PCA

The PCA+combination method estimates the significance of features by combining the results of different algorithms (GB, MI, and RFE with Lasso) and applying the PCA. The projection of normalized significances onto the principal component space is performed using the component matrix W obtained from PCA (11):

$$I_{PCA}(k) = \sum_{i=1}^{n} I_{normalized}(z_i) * |W_{k,i}| \quad (11)$$

where $I_{PCA}(k)$ is the significance of the k-th principal component. $W_{k,i}$ is an element of the PCA component matrix responsible for the contribution of the feature z_i to the k-th principal component.

Inverse transformation to the original space. The inverse transformation of the significances from the principal component space to the feature space is performed using the transposed component matrix W^T (12):

$$I_{original}(z_i) = \sum_{k=1}^{n} I_{PCA}(k) * |W_{k,i}| \quad (12)$$

Normalization of significances. The obtained values are normalized to ensure interpretability (their sum is 100%) (13):

$$I_{PCA}(z_i) = \frac{I_{original}(z_i)}{\sum_{j=1}^{n} I_{original}(z_j)} * 100 \quad (13)$$

where $I_{PCA}(z_i)$ is the final significance of the feature z_i, calculated after the inverse transformation from the principal component space.

f. SHAP analysis

SHAP analysis is used to estimate the contribution of each feature to the model prediction based on a proxy target variable. The proxy target variable (y_{proxy}) is formed through a linear combination of feature weights calculated using GB, MI, and RFE with Lasso.

Combined significance of features. The significance of each feature is calculated as an average value using three methods (14):

$$NormalizedImportance_i = \frac{GB_i + MI_i + RFE_i}{\sum_{j=1}^{N}(GB_j + MI_j + RFE_j)} \quad (14)$$

where GB_i is the significance of feature *i* obtained from GB, MI_i is the significance of feature *i*, calculated using MI, RFE_i is the significance of feature *i*, determined by the RFE method using Lasso, *N* is number of features, and $NormalizedImportance_i$ is the final normalized significance of feature *i*.

Creating a proxy target variable. The proxy target variable y_{proxy} is formed as a linear combination of features X using the normalized significance of the features (15):

$$y_{proxy} = \sum_{i=1}^{N} NormalizedImportance_i * x_i \quad (15)$$

where y_{proxy} is proxy target variable and x_i is value of feature *i*.

SHAP values for features. SHAP values are calculated based on the Shapley theory: SHAP estimates the contribution of each feature i for observation k through SHAP values (ϕ_i^k) which are calculated using the formula from the Shapley theory (16):

$$\phi_i^k = \sum_{S \subseteq N\{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (16)$$

where *S* is subset of features not including *i*, *N* is set of all signs, $f(S)$ is prediction of the f_GB model trained only on the feature subset *S*, and $f(S \cup \{i\})$ is prediction of the f_GB model trained on the feature subset S with the addition of feature i.

The combined importance of features ($NormalizedImportance_i$) determines their influence on the *formation* of the proxy target, which is created through a linear combination of features taking into account their relative importance. The GB model is used to account for complex relationships between features, allowing for a more accurate prediction of the proxy target variable. Additionally, the SHAP method is used to estimate the contribution of each feature to the model predictions, providing an interpretable importance. This approach combines information from different evaluation methods, which allows for complex interactions of features to be taken into account, and a more interpretable and accurate assessment of their contribution can be obtained.

g. Visualizing the results

For each target variable y, the results of I_{PCA} , $I_{combined}$, $and\ I_{SHAP}$ are normalized and visualized for the top 5 features (17):

$$I_{norm}(x_i) = \frac{I(x_i)}{\sum_{j=1}^{n} I(x_i)} * 100 \quad (17)$$

The final visualization is constructed as horizontal columns indicating each method's percentage values of importance. The methodological design adopted in this study aligns with recent advancements in intelligent agricultural systems and data-driven modeling approaches. Contemporary research highlights the growing role of AI-enabled architectures, IoT-supported monitoring tools, and machine learning pipelines in transforming agricultural decision-making and improving the efficiency of resource management [23]. Furthermore, modern intelligent machines and computational frameworks demonstrate that the integration of analytical models with real-time data streams significantly enhances the accuracy and responsiveness of agricultural analytics [24]. In addition, open-source AI architectures and MLOps-based workflows provide a scalable foundation for combining multiple analytical techniques, ensuring reproducibility, transparency, and operational adaptability in complex agricultural environments [25]. Building upon these developments, the proposed methodology offers a unified and extensible framework for robust feature evaluation and interpretable data analysis.

3. RESULTS AND DISCUSSION

After performing feature importance analysis using different methods, including PCA, SHAP, and combination approaches, we can visualize the results to understand their contribution to the target variables better. Figure 2 shows the results of feature importance analysis for the target variable “Market Capacity (%)” performed using three approaches: PCA+combination, simple combination, and SHAP analysis. Each method evaluates the contribution of features to the prediction of the target variable using its unique algorithms. The horizontal bar chart displays the five most significant features of the selected target variable. Each bar in the chart represents the relative importance of the feature expressed as a percentage. The three approaches are highlighted in different colors: blue for PCA+combination, orange for simple combination, and green for SHAP analysis.

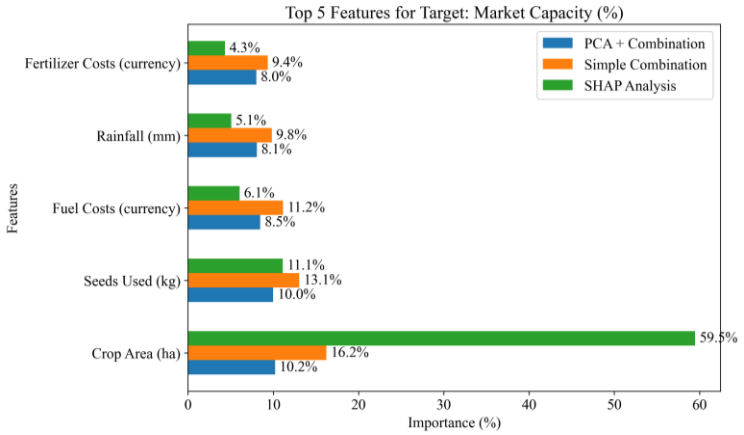

Figure 2. Comparison of the significance of features for “Market Capacity (%)” using PCA, simple combination, and SHAP methods

SHAP analysis highlights “Area of farmland (ha)” as the most significant feature (59.5%) due to its impact on production volume and market capacity, and also emphasizes the importance of such factors as “Number of seeds used (kg)” (11.1%) and “Fuel costs” (6.1%). A simple combination of methods showed a more even distribution of importance, with “Area of farmland” (16.2%) and “Number of seeds” (13.1%) remaining the key features, smoothing out local effects. PCA+combination focuses on features that explain the variance, with “Area of farmland” (10.2%) and “Number of seeds” (10.0%) remaining the leaders. The methods complement each other, providing a balanced and detailed analysis of the factors influencing market capacity. Figure 3 shows the assessment results of the significance of features for the target variable “Scenario Budget (Currency)”, calculated based on three approaches: PCA+combination, simple combination and SHAP analysis. Each method identifies the most significant factors influencing budget indicators.

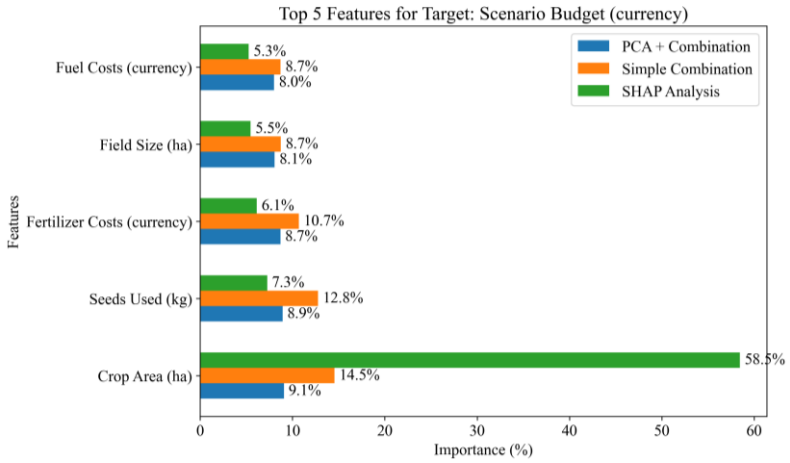

Figure 3. Comparison of the importance of features for "Scenario budget (currency)" using PCA, simple combination and SHAP methods

SHAP analysis showed that “Area of agricultural land (ha)” exerts the strongest influence on budget formation (58.5%), which is expected given its direct relationship with expenses for soil preparation, fertilization, and machinery operation. The method also revealed notable contributions from the “Number of seeds used” (7.3%) and “Fertilizer costs” (6.1%), emphasizing the role of input intensity in shaping budget outcomes. In contrast, the simple combination approach produced a more uniform distribution of factor weights: land area accounted for 14.5%, seed quantity for 12.8%, and fertilizer expenses for 10.7%. This distribution reflects a more balanced view of the underlying relationships and reduces the dominance of any single variable. The PCA+combination method, which captures global variance structure, similarly positioned land area (9.1%) and seed quantity (8.9%) as the most influential variables, but it further attenuated the contributions of secondary features due to its focus on overall data dispersion. Taken together,

the three approaches complement each other and offer a multidimensional understanding of the determinants shaping the scenario budget, integrating nonlinear effects, structural balance, and variance-driven trends. Figure 4 summarizes the assessment of feature importance for the target variable "Scenario Rating (Points)", illustrating how each analytical method PCA+combination, simple combination, and SHAP provides a distinct yet coherent interpretation of the relative influence of individual indicators on the final rating.

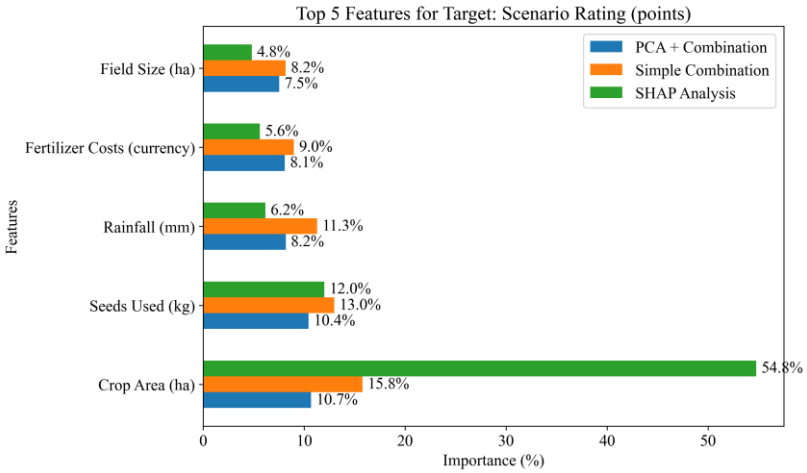

Figure 4. Comparison of the importance of features for "Scenario Rating (scores)" using PCA, simple combination, and SHAP methods

SHAP analysis showed that "Crop area (ha)" is the most influential determinant of the scenario rating (54.8%). This strong effect reflects the fundamental role of cultivated land size in shaping productivity potential, operational efficiency, and overall performance outcomes. SHAP also indicated meaningful contributions from the "Number of seeds used" (12.0%) and "Rainfall" (6.2%), underscoring the combined impact of input intensity and climatic conditions on scenario evaluation. In contrast, the simple combination method produced a more moderate distribution of weights: crop area remained the leading factor (15.8%), followed closely by seed use (13.0%), suggesting a more balanced representation of input-driven relationships. The PCA+combination approach similarly identified crop area (10.7%) and seed quantity (10.4%) as dominant contributors, but its variance-oriented nature reduced the prominence of secondary variables and mitigated collinearity effects. Together, these analytical perspectives offer a multidimensional view of the drivers influencing scenario ranking, capturing both localized nonlinear interactions and broader structural patterns within the dataset. Figure 5 illustrates the subsequent assessment of feature importance for the target variable "Yield Increase (%)" obtained using the three analytical frameworks: PCA+combination, simple combination, and SHAP.

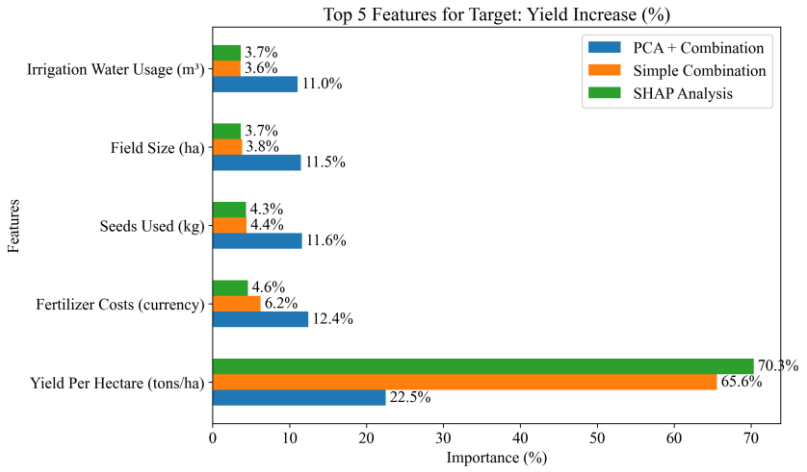

Figure 5. Feature importance for yield increase (%) across three analytical methods

SHAP analysis demonstrated that "Yield per hectare (tonnes/ha)" is the dominant contributor to yield increase (70.3%), which is expected given that this indicator reflects the combined effects of soil quality, crop management practices, and climatic conditions. Although "Fertilizer Cost" (4.6%) and "Seed Number" (4.3%) exhibited smaller individual effects, they still represent important components of input efficiency and nutrient availability. The simple combination approach produced a more moderate distribution of influences, again confirming the central role of yield per hectare (65.6%), while elevating the contributions of supporting factors such as fertilizer expenses (6.2%). Unlike the previous methods, PCA+combination identified broader structural patterns in the data: yield per hectare remained the most impactful feature (22.5%), but additional variables—including field size (11.5%) and water usage (11.0%)—emerged as influential due to their significance within global variance components. Collectively, these perspectives highlight both immediate agronomic drivers and underlying systemic dependencies, offering a richer interpretation of factors shaping yield enhancement. Figure 6 further examines the relative importance of variables associated with cost reduction (%), comparing the outcomes derived from PCA+combination, simple combination, and SHAP analysis.

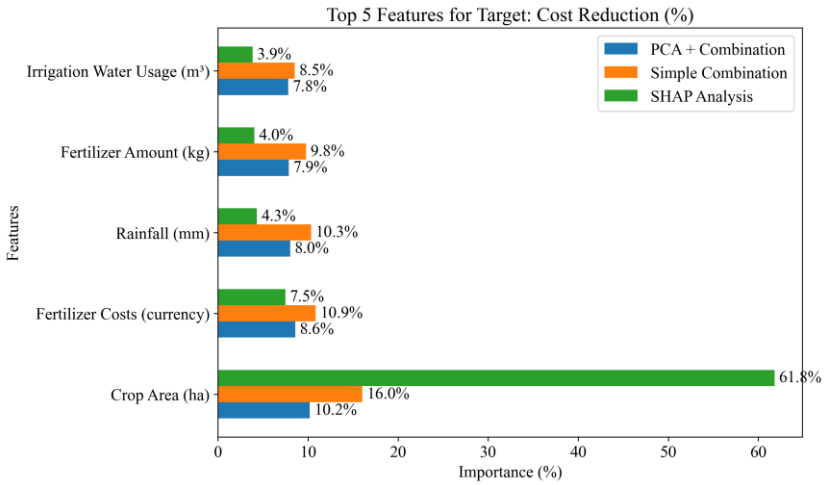

Figure 6. Impact of attributes on cost reduction (%)

SHAP analysis showed that the main cost reduction factor was "Land area (ha)" (61.8%), optimizing costs through economies of scale, and also highlighted "Fertilizer cost" (7.5%) and "Rainfall" (4.3%) as significant features. A simple combination of methods distributed the significance more evenly, confirming the importance of "Land area" (16.0%) and increasing the contribution of such factors as "Fertilizer cost" (10.9%) and "Rainfall" (10.3%). PCA+combination highlighted "Land area" (10.2%) and "Rainfall" (8.0%) through the analysis of global relationships. Using all methods together allows us to consider complex nonlinear relationships, identify general patterns, and build a balanced analysis, which is essential for cost management strategies and improving the efficiency of agricultural production. Figure 7 presents the analysis results of the significance of factors influencing risk reduction (%) using the PCA+combination, simple combination, and SHAP analysis methods.

SHAP analysis indicated that "Distance to Market (km)" is the most influential determinant of risk reduction (68.4%), reflecting the substantial effect of transportation distance on logistics reliability, product preservation, and overall operational vulnerability. The method also underscored the roles of "Transportation Cost" (12.4%) and "Fertilizer Cost" (4.1%), suggesting that financial and input-related factors contribute to risk exposure, albeit to a lesser extent. Results from the simple combination approach aligned with this pattern: distance to market remained the dominant variable (61.7%), while transportation cost (13.6%) maintained a significant secondary influence, reinforcing the connection between logistical constraints and risk outcomes. PCA+combination, which emphasizes structural variance, likewise positioned distance to market (28.7%) as the leading factor but provided additional perspective by elevating the importance of transportation and fertilizer costs (10.0% and 8.1%, respectively) within global data relationships. Collectively, these complementary methods reveal that logistical accessibility is a central driver of risk mitigation, while economic factors shape the broader risk environment. Figure 8 presents the subsequent

evaluation of feature importance for capacity utilization (%), comparing insights derived from PCA+combination, simple combination, and SHAP analysis.

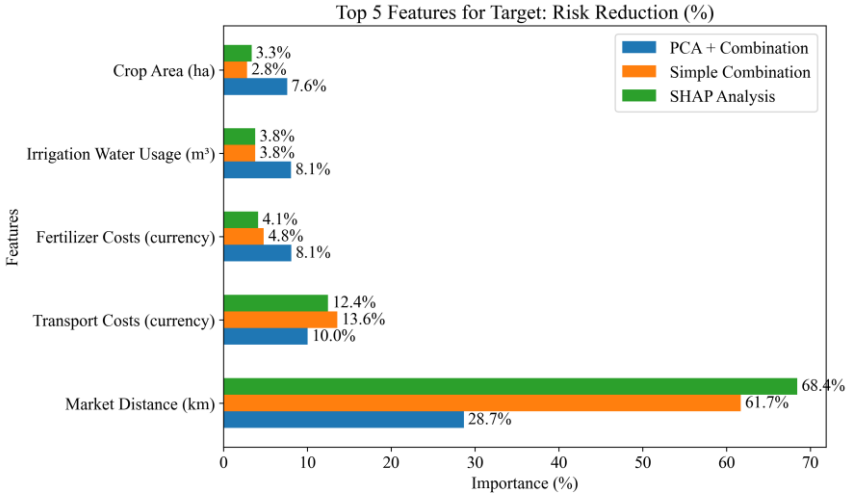

Figure 7. Impact of signs on risk reduction (%)

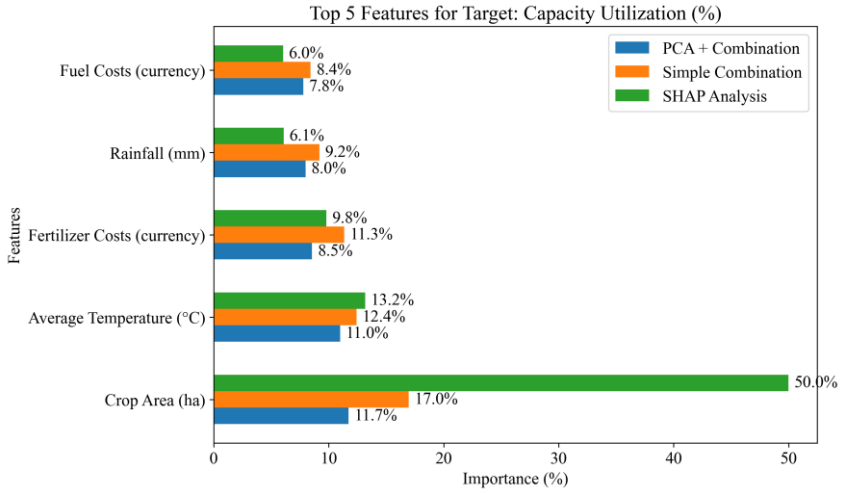

Figure 8. Impact of attributes on capacity utilization (%)

SHAP analysis showed that “Land Area (ha)” exerts the strongest influence on capacity utilization (50.0%), underscoring how production scale shapes the ability to fully leverage available resources. The analysis also revealed notable contributions from “Average Temperature” (13.2%) and “Fertilizer Cost” (9.8%), indicating that climatic conditions and input intensity meaningfully affect operational efficiency. The simple combination method produced a similar hierarchy of factors, placing land area at the forefront (17.0%) while presenting a more balanced distribution for average temperature (12.4%) and fertilizer expenses (11.3%). In contrast, PCA+combination emphasized a more evenly structured pattern of influence, identifying land area (11.7%) and average temperature (11.0%) as principal contributors within the broader variance framework. Together, these approaches provide a multi-layered view of capacity utilization dynamics by capturing both immediate, localized interactions and overarching systemic relationships—an essential foundation for informed strategic planning and resource allocation. Figure 9 further explores the significance of features shaping seasonal profit, integrating insights generated through PCA+combination, simple combination, and SHAP analysis to determine the primary economic drivers of profitability.

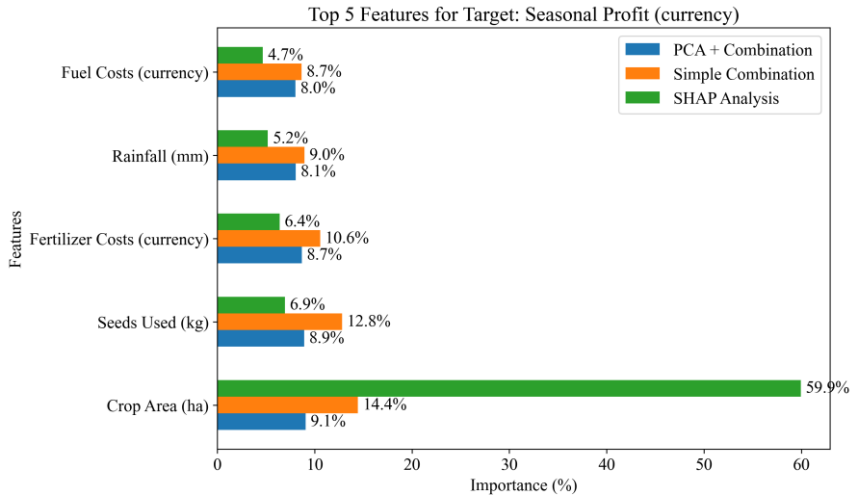

Figure 9. Analysis of factors affecting seasonal profit (currency)

SHAP analysis showed that "Land Area (ha)" is the key factor for increasing seasonal profit (59.9%), while "Seed Use (6.9%)" and "Fertilizer Cost (6.4%)" play a significant role in resource optimization. Simple combination confirmed the importance of "Land Area" (14.4%) and highlighted the contribution of "Seed Use" (12.8%) and "Fertilizer Cost" (10.6%), providing a balanced analysis. PCA+combination distributed the importance evenly, with an emphasis on "Land Area" (9.1%), "Seed Use" (8.9%), and "Fertilizer Cost" (8.7%). The combined use of the methods provides a comprehensive understanding of the factors affecting profit, which is essential for strategic management and optimization. Figure 10 presents the analysis of the significance of factors influencing the actual sales volume, performed using three approaches: PCA+combination, simple combination, and SHAP analysis. These methods allowed us to identify the main features and relationships of the factors determining the results.

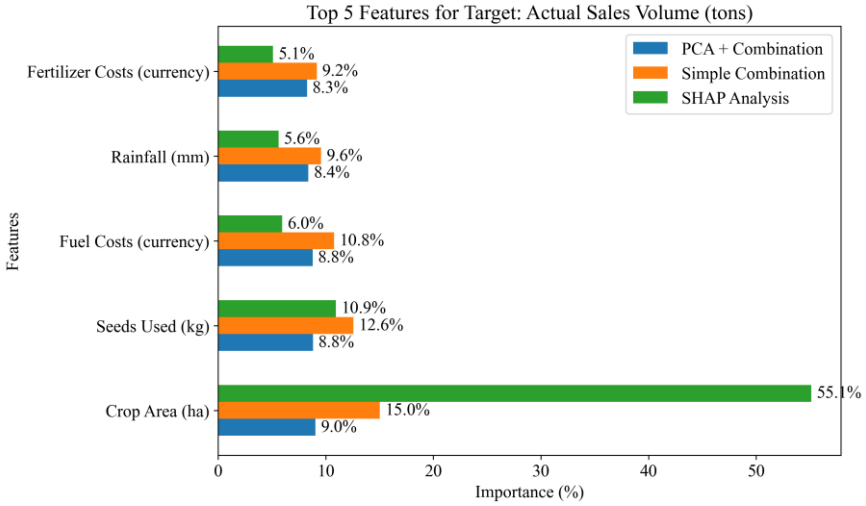

Figure 10. Impact of factors on actual sales volume (tons)

SHAP analysis revealed that "Crop Area (ha)" plays the dominant role in determining actual sales volume (55.1%), reflecting the intuitive link between cultivated land size and production output. The model also attributed meaningful influence to "Seed Use" (10.9%), as well as operational and climatic variables such as "Fuel Cost" (6.0%) and "Rainfall" (5.6%), which collectively affect harvesting capacity and crop performance. The simple combination method supported these patterns, again positioning crop area (15.0%) and seed use (12.6%) as leading contributors, while assigning a stronger weight to fuel cost (10.8%), emphasizing the relevance of energy expenditures in production and transportation processes. Meanwhile, the PCA+combination approach, which captures overarching variance structure, presented a more evenly spaced

distribution of influences, with crop area (9.0%), seed use (8.8%), and fertilizer cost (8.3%) emerging as the most impactful variables within global data patterns. Taken together, the results demonstrate the interconnected roles of production scale, input utilization, and operational conditions, offering a robust foundation for improving sales forecasting and strategic decision-making in agricultural systems. Figure 11 extends this analysis to sales revenue, comparing how the three analytical frameworks—PCA+combination, simple combination, and SHAP—characterize the relative contribution of key determinants and provide complementary perspectives on revenue formation.

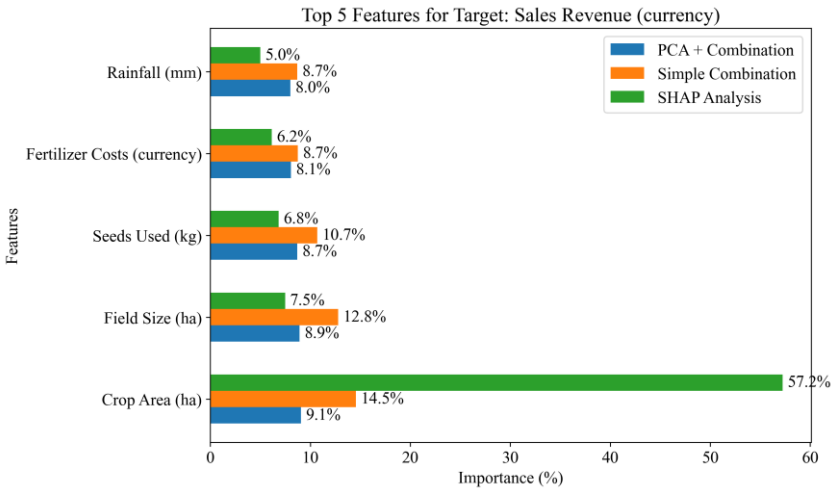

Figure 11. Factors' importance for the target variable: sales revenue (currency)

SHAP analysis identified Land Area (ha) as the key revenue driver (57.2%), while Seed Use (6.8%) and fertilizer cost (6.2%) highlighted the importance of resource optimization. Simple combination confirmed the importance of Land Area (14.5%), adding the contribution of Seed Use (10.7%) and fertilizer cost (8.7%), providing a balanced analysis. PCA+combination distributed the importance equally, highlighting Land Area (9.1%) and Seed Use (8.7%) as the key drivers. The combined application of the methods considers both non-linear relationships and global trends, providing a comprehensive understanding of the factors affecting revenue, which is essential for strategic management in the agribusiness sector. Based on cooperative game theory, SHAP analysis provided a quantitative assessment of the contribution of features to the target variables, accounting for nonlinear dependencies and interactions. Using proxy target variables, SHAP identified “Land area (ha)” as a key driver for “Market capacity” (59.5%) and “Sales revenue” (57.2%), highlighting the importance of production scale. The method also revealed local effects, such as the influence of logistic factors such as “Distance to market” (68.4%) on risk mitigation. SHAP has proven its effectiveness in problems requiring profound interpretation of complex relationships. Combining GB, MI, and RFE with Lasso provides a balanced analysis of factors, accounting for linear and nonlinear relationships. For scenario budget, the key factors were land area (14.5%), seed use (12.8%), and fertilizer costs (10.7%), reflecting their importance in operating expenses. PCA+combination identifies global trends and interdependent data structures, focusing on the features that introduce the most significant variance. For example, for yield increase, the leaders were Yield per Hectare (22.5%) and Field Size (11.5%). Both methods effectively analyze local and global dependencies, which is essential for strategic planning.

Despite the robustness of the multi-method framework, several limitations should be noted. The analysis is constrained by the available dataset, which may not fully capture the diversity of agronomic, climatic, and economic variables relevant to broader agricultural systems. The applied methods rely on static, historical data and therefore do not reflect temporal variability or seasonal fluctuations. Additionally, the use of a proxy target variable for SHAP interpretation introduces a degree of simplification in modeling complex non-linear relationships. These limitations highlight the need for future research incorporating temporal models, richer datasets, and cross-regional validation.

4. CONCLUSION

This study set out to develop a comprehensive methodological framework for evaluating feature importance in agricultural process management by integrating three analytical approaches: SHAP analysis,

simple combination, and PCA+combination. The overarching objective was to create an interpretable, robust, and multi-perspective tool capable of identifying the most influential factors that shape agricultural productivity, efficiency, and economic performance. The results of the study demonstrate that the combined use of complementary analytical techniques offers a significantly richer understanding of factor importance compared to relying on individual methods. While SHAP provides local interpretability and captures nonlinear interactions, the simple combination approach offers balanced feature ranking, and PCA+combination highlights global variance-driven structures. Together, these perspectives enable a more holistic assessment of the factors that influence key agricultural indicators. This integrative framework supports informed and transparent decision-making processes and improves the interpretability of complex agricultural data systems.

The main contribution of this research lies in the development of a unified analytical methodology that can be applied across various agricultural scenarios to identify and prioritize critical factors affecting resource allocation, operational efficiency, and strategic planning. By offering a methodology that accommodates both nonlinear dependencies and global structural patterns, the study provides a valuable tool for researchers, policymakers, and agribusiness practitioners seeking to enhance production management and optimize economic outcomes. Looking ahead, several directions for future research are proposed. First, expanding the dataset with additional agro-climatic, soil, and economic variables would further enhance the robustness of the model. Second, incorporating temporal and seasonal dynamics through time-series models or recurrent neural architectures may improve the system's ability to capture fluctuations inherent to agricultural processes. Third, validating the methodology across different geographic regions and crop types would strengthen its generalizability. Finally, integrating the developed approach into real-time decision-support platforms could significantly enhance its practical applicability within precision agriculture. Overall, the study provides a scientifically grounded and practically relevant framework that advances the understanding of factor importance in agricultural systems and opens new avenues for the development of data-driven management strategies.

FUNDING INFORMATION

This research received no external funding.

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
Gulzira Abdikerimova	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Moldir Yessenova	✓		✓		✓		✓	✓		✓			✓	✓
Aizhan Zharkimbekova	✓		✓	✓			✓			✓	✓		✓	✓
Zhanar Beldeubayeva	✓		✓	✓			✓			✓	✓		✓	✓
Aigulim Bayegizova		✓				✓		✓	✓	✓	✓	✓		
Nurgul Uzakkyzy		✓			✓		✓			✓		✓		✓
Ainagul Alimagambetova		✓				✓		✓	✓	✓	✓	✓		
Gulden Murzabekova	✓		✓	✓			✓			✓	✓		✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] R. Raman, H. Kantari, A. A. Gokhale, K. Elangovan, B. Meenakshi, and S. Srinivasan, "Agriculture Yield Estimation Using Machine Learning Algorithms," in *2024 International Conference on Automation and Computation (AUTOCOM)*, Dehradun, India, 2024, pp. 187–191, doi: 10.1109/AUTOCOM60220.2024.10486107.

[2] O. B. Akintuyi, "Adaptive AI in precision agriculture: A review: Investigating the use of self-learning algorithms in optimizing farm operations based on real-time data," *Open Access Research Journal of Multidisciplinary Studies*, vol. 7, no. 2, pp. 16–30, 2024, doi: 10.53022/oarjms.2024.7.2.0023.

[3] D. Huo, A. W. Malik, S. D. Ravana, A. U. Rahman, and I. Ahmedy, "Mapping smart farming: Addressing agricultural challenges in data-driven era," *Renewable and Sustainable Energy Reviews*, vol. 189, p. 113838, 2024, doi: 10.1016/j.rser.2023.113858.

[4] 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.

[5] J. Tussupov *et al.*, "Analyzing Disease and Pest Dynamics in Steppe Crop Using Structured Data," *IEEE Access*, vol. 12, pp. 71323–71330, 2024, doi: 10.1109/ACCESS.2024.3397843.

[6] E. B. Wijayanti, D. R. I. M. Setiadi, and B. H. Setyoko, "Dataset Analysis and Feature Characteristics to Predict Rice Production based on eXtreme Gradient Boosting," *Journal of Computing Theories and Applications*, vol. 1, no. 3, pp. 299–310, 2024, doi: 10.62411/jcta.10057.

[7] N. V. V. Reddy and T. Manimegalai, "Predicting the crop yield in agriculture using gradient boosting algorithm in comparison of naive bayes algorithm," *AIP Conference Proceedings*, vol. 2853, no. 1, 2024, doi: 10.1063/5.0204781.

[8] E. S. M. El-Kenawy, A. A. Alhussan, N. Khodadadi, S. Mirjalili, and M. M. Eid, "Predicting Potato Crop Yield with Machine Learning and Deep Learning for Sustainable Agriculture," *Potato Research*, vol. 68, no. 1, pp. 759–792, 2025, doi: 10.1007/s11540-024-09753-w.

[9] Y. Yang, X. Zhou, J. Xu, H. Wang, L. Liu, and W. Cao, "Coupling mutual information into ecological networks to analyze the sustainability of water-energy nexus: A case study of Yangtze River Economic Belt," *Journal of Cleaner Production*, vol. 448, p. 141705, 2024, doi: 10.1016/j.jclepro.2024.141705.

[10] X. Zhao, M. Zhang, Z. Xiao, and B. Kang, "Evaluating the reliability and relative weight of the evidence using approximate evidential mutual information," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108409, 2024, doi: 10.1016/j.engappai.2024.108409.

[11] V. V. Kamaraj, P. S. Maran, and P. Ranjana, "Modified Recursive Feature Elimination (MRFE) Technique for Soil Classification and Better Crop Production," *AIP Conference Proceedings*, vol. 3075, no. 1, 2024, doi: 10.1063/5.0217607.

[12] J. A. Ingio, A. S. Nsang, and A. Iorliam, "Optimizing Rice Production Forecasting Through Integrating Multiple Linear Regression with Recursive Feature Elimination," *Journal of Future Artificial Intelligence and Technologies*, vol. 1, no. 2, pp. 96–108, 2024, doi: 10.62411/faith.2024-17.

[13] L. Xiong *et al.*, "Improved support vector regression recursive feature elimination based on intragroup representative feature sampling (IRFS-SVR-RFE) for processing correlated gas sensor data," *Sensors and Actuators B: Chemical*, vol. 419, p. 136395, 2024, doi: 10.1016/j.snb.2024.136395.

[14] A. R. Barzani, P. Pahlavani, O. Ghorbanzadeh, K. Gholamnia, and P. Ghamisi, "Evaluating the Impact of Recursive Feature Elimination on Machine Learning Models for Predicting Forest Fire-Prone Zones," *Fire*, vol. 7, no. 12, p. 440, 2024, doi: 10.3390/fire7120440.

[15] D. Hammoumi *et al.*, "Seasonal Variations and Assessment of Surface Water Quality Using Water Quality Index (WQI) and Principal Component Analysis (PCA): A Case Study," *Sustainability (Switzerland)*, vol. 16, no. 13, p. 5644, 2024, doi: 10.3390/su16135644.

[16] O. T. Faloye, A. E. Ajayi, V. Kamchoom, O. A. Akintola, and P. G. Oguntunde, "Evaluating Impacts of Biochar and Inorganic Fertilizer Applications on Soil Quality and Maize Yield Using Principal Component Analysis," *Agronomy*, vol. 14, no. 8, p. 1761, 2024, doi: 10.3390/agronomy14081761.

[17] W. Wei, P. Yan, L. Zhou, H. Zhang, B. Xie, and J. Zhou, "A comprehensive drought index based on spatial principal component analysis and its application in northern China," *Environmental Monitoring and Assessment*, vol. 196, no. 2, p. 193, 2024, doi: 10.1007/s10661-024-12366-y.

[18] H. Jin *et al.*, "Spatiotemporal evolution of drought status and its driving factors attribution in China," *Science of the Total Environment*, vol. 958, p. 178131, 2025, doi: 10.1016/j.scitotenv.2024.178131.

[19] Y. Wang, P. Wang, K. Tansey, J. Liu, B. Delaney, and W. Quan, "An interpretable approach combining Shapley additive explanations and LightGBM based on data augmentation for improving wheat yield estimates," *Computers and Electronics in Agriculture*, vol. 229, p. 109758, 2025, doi: 10.1016/j.compag.2024.109758.

[20] S. S. L. D. Arumugam, R. P. Kumar, A. Rubeshkumar, and S. S. Rahul, "Predicting Crop Yield using Long Short-Term Memory, Integrated Gradients and Shapley Additive Explanations," in *2024 4th International Conference on Pervasive Computing and Social Networking (ICPCSN)*, Salem, India, 2024, pp. 975–983, doi: 10.1109/ICPCSN62568.2024.00163.

[21] A. Aldrees, M. Khan, A. T. B. Taha, and M. Ali, "Evaluation of water quality indexes with novel machine learning and SHapley Additive ExPlanation (SHAP) approaches," *Journal of Water Process Engineering*, vol. 58, p. 104789, 2024, doi: 10.1016/j.jwpe.2024.104789.

[22] K. D. Bissadu, S. Sonko, and G. Hossain, "Society 5.0 enabled agriculture: Drivers, enabling technologies, architectures, opportunities, and challenges," *Information Processing in Agriculture*, vol. 12, no. 1, pp. 112–124, 2025, doi: 10.1016/j.inpa.2024.04.003.

[23] K. Pachiappan, K. Anitha, R. Pitchai, S. Sangeetha, T. V. V. Satyanarayana, and S. Boopathi, "Intelligent machines, IoT, and AI in revolutionizing agriculture for water processing," in *Handbook of Research on AI and ML for Intelligent Machines and Systems*, IGI Global Scientific Publishing, 2023, pp. 374–399, doi: 10.4018/978-1-6684-9999-3.ch015.

[24] N. Victor *et al.*, "Remote Sensing for Agriculture in the Era of Industry 5.0-A Survey," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 5920–5945, 2024, doi: 10.1109/JSTARS.2024.3370508.

[25] A. C. Cob-Parro, Y. Lalangui, and R. Lazcano, "Fostering Agricultural Transformation through AI: An Open-Source AI Architecture Exploiting the MLOps Paradigm," *Agronomy*, vol. 14, no. 2, p. 259, 2024, doi: 10.3390/agronomy14020259.

BIOGRAPHIES OF AUTHORS

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 at 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: abdikerimova_gb@enu.kz.

Moldir Yessenova received a bachelor's degree in information systems in 2014 and a master's degree in 2017 in information technology from L.N. Gumilyov Eurasian National University, Kazakhstan, Nursultan. Currently, she is a doctoral student at the Department of Information Systems at the L.N. Gumilyov Eurasian National University. Her research interests include image processing, computer vision, satellite imagery, artificial intelligence, and machine learning. She can be contacted at email: moldirrespect@gmail.com.

Aizhan Zharkimbekova Ph.D. in the specialty "Computer Science". Currently, she works at the NJSC L. N. Gumilyov Eurasian National University, Department of Information Security, as a Senior Lecturer. She has more than 22 years of scientific and pedagogical experience and has authored over 50 scientific papers, including 4 articles indexed in the Scopus database and 15 teaching aids. Her Hirsch index is 3. Her research interests include information and communication technologies, internet technologies, cloud technologies, information security, computer network security, and the mathematical foundations of information protection. She can be contacted at email: zh.aizhan.t@gmail.com.

Zhanar Beldeubayeva received her Ph.D. in 2019 in Information Systems from Serikbayev East Kazakhstan Technical University, Kazakhstan. Currently, she is a senior lecturer at the Department of Information Systems, Kazakh Agrotechnical Research University named after S. Seifullin. Her research interests include process modeling, knowledge management, data mining, and machine learning. She can be contacted at email: zbeldeubayeva@list.ru.

Aigulim Bayegizova graduated from S.M. Kirov Kazakh State University in 1982 with a degree in Applied Mathematics. In 2010, she defended her Ph.D. thesis in the specialty "01.01.02 – differential equations and mathematical physics" and received the degree of Candidate of Physical and Mathematical Sciences. She began her career in 1982 as an assistant at the Department of Higher Mathematics of the Dzhezkazgan branch of the Karaganda Polytechnic Institute. Currently, she is a senior lecturer at the Department of Radio Engineering, Electronics and Telecommunications of the L.N. Gumilyov Eurasian National University. She is the author of more than 60 scientific papers, including 1 monograph, 6 articles in the Scopus database. Research interests – programming, information security and information protection, artificial intelligence, and cloud technologies. She can be contacted at email: baegiz_a@mail.ru.

Nurgul Uzakkyzy received her Ph.D. in 2020 in Computer Engineering and Software from L.N. Gumilyov Eurasian National University, Kazakhstan. Currently. She is an associate professor of the Department of Computer and Software Engineering at the same university. Her research interests include signal processing, image processing, computer vision, satellite imagery, artificial intelligence, and machine learning. She can be contacted at email: nura_astana@mail.ru.

Ainagul Alimagambetova defended her dissertation for the degree of candidate of physical and mathematical sciences in 2009 at the L.N. Gumilyov Eurasian National University, Kazakhstan. Currently, she works as a senior lecturer at the Department of Information Systems at the L.N. Gumilyov Eurasian National University. Her research interests include mathematical modeling, mathematical foundations of cryptography, and new information technologies. She can be contacted at email: ainash_777@mail.ru.

Gulden Murzabekova graduated from the Faculty of Applied Mathematics-Control Processes of Saint-Petersburg State University in 1994, where she also successfully defended her doctoral thesis in 1997 on discrete mathematics and mathematical cybernetics. Since 1998, she has been an associate professor in the Department of Informatics, serving as head of the Department of Information and Communication Technologies from 2003 to 2022. She is currently an assistant professor in the Department of Computer Science at Seifullin Kazakh Agrotechnical University. She has authored more than 100 papers. Her research interests include numerical methods of nonsmooth analysis and nondifferentiable optimization, mathematical modeling, artificial intelligence, and machine learning. She can be contacted at email: g.murzabekova@kazatu.kz.