International Journal of
Agriculture, Environment and Food Sciences
e-ISSN: 2618-5946 https://dergipark.org.tr/jaefs

DOI: https://doi.org/10.31015/2025.si.5 Int. J. Agric. Environ. Food Sci. 2025; 9 (Special Issue): xxx–xxx

Exploratory Data Analysis and Kernel Feature Fusion for Enhanced SVM and Random Forest–Based Crop Recommendation

Kutalmis TURHAL[1], **Umit Cigdem TURHAL**[2]

[1]Biosystem Engineering, Faculty of Agriculture and Natural Sciences, Bilecik Seyh Edebali University, Bilecik, Türkiye
[2]Electrical Electronics Engineering, Faculty of Engineering, Bilecik Seyh Edebali University, Bilecik, Türkiye

Article History
Received: July 31, 2025
Accepted: September 12, 2025
Published Online: November 13, 2025

Article Info
Type: Research Article
Subject: Sustainable Agricultural Development

Corresponding Author
Ümit Çiğdem Turhal
ucigdem.turhal@bilecik.edu.tr

Author ORCID
[1]https://orcid.org/0000-0002-5347-8513
[2]https://orcid.org/0000-0003-2387-1637

Available at
https://dergipark.org.tr/jaefs/issue/93587/1754586

Abstract

Modern crop recommendation systems must accurately grasp the complex and nonlinear relationships between soil nutrients to support effective agricultural decisions. In this study, we introduce a framework that combines supervised and unsupervised learning through kernel feature fusion, integrating Radial Basis Function (RBF) Kernel Principal Component Analysis (KPCA) and Kernel Linear Discriminant Analysis (KLDA) into a single seven-dimensional embedding. First, six principal components are extracted using RBF-KPCA to capture global nonlinear variance in the raw data. Similarly, in the raw space, an Nystroem-approximated RBF transformation followed by LDA produces a single discriminant axis (KLDA) for better supervised class separation. These features are fused by concatenation and then input into Support Vector Machine (SVM) classifiers (using polynomial and RBF kernels) and a Random Forest (RF) classifier. In the experiments, a publicly available dataset comparing maize and barley based on six soil features was used. The fused representation significantly outperformed raw data and single-embedding methods, with Polynomial SVM increasing by 18.5%, RBF SVM improving by 10.1%, and RF rising by 4.7% over the raw data. These results show that combining unsupervised variance maximization with supervised discriminant projection creates a richer, more discriminative feature space—especially beneficial for SVMs in crop recommendation tasks. Our kernel fusion approach offers a powerful and flexible strategy for precision agriculture, enabling robust decision support without extensive field trials or repeated laboratory tests.

Keywords: Precision Agriculture, Crop Recommendation, Soil Nutrient Embedding, Kernel Feature Fusion, Artificial Intelligence

Cite this article as: Turhal, K., Turhal, U.C. (2025). Exploratory Data Analysis and Kernel Feature Fusion for Enhanced SVM and Random Forest–Based Crop Recommendation. International Journal of Agriculture, Environment and Food Sciences, 9 (Special Issue): xxx–xxx. https://doi.org/10.31015/2025.si.5

INTRODUCTION

As the global population continues to grow, and in light of climate variability and limited agricultural resources, optimizing crop selection has become essential for modern agriculture (Gunawan et al., 2024; Liakos et al.,2018). Recommending suitable crops for cultivation based on environmental and soil conditions not only boosts agricultural productivity but also supports sustainable land use and efficient resource management. Historically, crop recommendations relied heavily on expert knowledge, historical data, and trial-and-error methods, which often fail to adapt to rapidly changing climate patterns and regional soil differences. Recent advancements in Artificial Intelligence (AI) have created new opportunities for data-driven decision-making in agriculture (Mishra and Mishra, 2024; Linaza et al., 2021). An analysis of OECD AI investment data for the G7 and Turkey shows that AI investments in agriculture are generally lower than in other sectors. However, Turkey and Canada emerge as two of the leading investors in agricultural AI (Çağlar, 2024).

Effective crop recommendations rely on identifying meaningful patterns from intricate soil and environmental data—this challenge is particularly evident in remote sensing and precision agriculture (Getahun et al., 2024; Bandara et al., 2020; Gosai et al., 2021). This process is called Feature Extraction and it is a crucial step in developing AI models, particularly when working with high-dimensional and diverse datasets like those found in agriculture (Ruano-Ordás, 2024). This process involves converting raw data into a set of informative, non-redundant features that effectively represent the underlying structure of the data (Zebari et al., 2020). By doing so, it enhances model accuracy, generalizability, and computational efficiency (Cheng, 2024).

Feature extraction is an essential step in AI pipelines and is frequently utilized in precision agriculture applications, including production forecasting (Gür, 2024) and robotic harvesting (Kahya & Aslan, 2024). This process is vital because extracting informative features greatly improves model performance and informs better decision-making (Ruano-Ordás, 2024). Feature extraction in agricultural decision-making systems, such as crop recommendation, serves two main purposes (Barburiceanu et al., 2021). Firstly, it allows for dimensionality reduction, which reduces noise, redundancy, and computational demands by decreasing the number of input variables while preserving the most relevant information. Secondly, it supports representation learning, enabling the extraction of high-level, abstract patterns from raw data—especially from temporal or structured inputs—that are often challenging to capture through manual feature engineering. However, the performance of these models heavily depends on the quality and structure of the input features, highlighting the importance of effective feature extraction (Taye, 2023).

Principal Component Analysis (PCA) is commonly used in agricultural studies, particularly in soil analyses, to reduce dimensionality, reveal relationships among variables, and visualize regional or spatial differences. The study performed by Yanardağ (2025) used the PCA algorithm to analyze the relationship between soil nutrients and to reduce the data size. The study highlighted that the principal components revealed regional differences, thus highlighting the need to prioritize organic matter management and pH regulation. However conventional machine-learning pipelines often depend on manually crafted features or linear dimensionality reduction techniques, such as PCA, which may not effectively capture the complex, nonlinear class boundaries present in soil-nutrient interactions (Kusuma et al., 2025). Kernel-based methods address this limitation by projecting data into high-dimensional feature spaces via a Mercer kernel, where classes that are inseparable in the original space become linearly separable (Anowar et al., 2021). KPCA enhances variance in an unsupervised manner by applying PCA to the kernel Gram matrix, effectively capturing global nonlinear structures (Briscik et al., 2023). In addition, KLDA focuses on supervised discriminant projection by first approximating the kernel using a Nystroem transform and then applying LDA to extract the axes that most effectively separate class means (Baudat & Anouar, 2000).

Feature fusion has emerged as a powerful technique in various machine learning domains, as it combines complementary representations to create more robust and discriminative feature sets. Early work by Bishop and Tipping (2000) demonstrated that concatenating features from multiple sources can enhance the performance of generative models. In recent years, various feature fusion strategies have been employed in the field of precision agriculture with distinct objectives (Upadhyay et al., 2025). These include crop classification, fertilizer recommendation, crop yield prediction, crop recommendation, pest detection, and others (Huang et al., 2024; Swaminathan et al., 2023; Boppudi and Jayachandran, 2024; Mahesh et al., 2024; Jiao et al., 2022).

In the context of kernel methods, a few studies arise on feature ranking fusion, crop classification. Especially fusion of KPCA and KLDA features used for tea classification in the study mentioned in Kaushal et al., (2022). Their results indicated that the nonlinear, kernel-based methods (KPCA and KLDA) outperformed the traditional linear approaches (PCA and LDA) in classification accuracy.

This study is the first, to our knowledge, to integrate Kernel Principal Component Analysis (KPCA) and Kernel Linear Discriminant Analysis (KLDA) into a dedicated, end-to-end feature extraction pipeline specifically designed for nutrient-based crop recommendations. Although previous research, such as that by Kaushal et al. (2022), has applied KPCA and KLDA separately in areas like tea classification, demonstrating that nonlinear kernel methods can outperform linear alternatives, no study has systematically evaluated the combination of KPCA and KLDA features for agricultural recommendation tasks. Our work addresses this gap by proposing a principled kernel fusion strategy and benchmarking it directly against single-embedding and raw-feature baselines.

We demonstrate that combining an unsupervised Kernel Principal Component Analysis (KPCA) embedding, which captures global nonlinear variance patterns, with a supervised Kernel Linear Discriminant Analysis (KLDA) discriminant axis, which focuses on class-specific information, retains complementary insights that either method alone might overlook. Through comprehensive experiments across various classifiers, including SVM variants and Random Forest, the fused representation consistently shows improvements in classification accuracy and the quality of decision boundaries. In addition to enhanced empirical performance, this fusion approach also increases robustness against noise and offers a practical pipeline (comprising kernel approximation, supervised projection, and fusion) that practitioners can easily adopt when working with small to moderately sized nonlinear agricultural datasets.

In this study, we developed a kernel feature fusion pipeline for nutrient-based crop recommendations. We began with a RF–based feature ranking on the raw six-dimensional nutrient dataset to identify the most informative variables: potassium (K), nitrogen (N), zinc (Zn), sulfur (S), soil pH, and phosphorus (P). The selected features were then mapped through two complementary kernel transformations: an RBF Kernel PCA (KPCA) that captures global nonlinear variance (retaining the top m components), Nystroem approximation followed by LDA (KLDA) that produces a supervised discriminant axis. The KPCA components and the single KLDA score were concatenated to form a compact (m+1)-dimensional fused embedding, which serves as input to downstream classifiers. By removing low-importance attributes before kernel embedding, noise is reduced, and the subsequent nonlinear projections focus on the variables most critical for distinguishing maize from barley.

To leverage the complementary strengths of unsupervised variance maximization and supervised discriminant projection (Kempfert et al., 2020; Peng & Zhao 2023), the six selected nutrients were processed through two parallel kernel mappings. The RBF-KPCA method extracted the top six nonlinear principal components, which capture the global variance structure, while a Nystroem approximation combined with LDA yielded a single supervised axis KLDA that was optimized for class separation.

These six KPCA components were concatenated with the one KLDA score to form a unified (m + 1)-dimensional feature vector. This feature fusion strategy allows the holistic nonlinear patterns uncovered by KPCA and the class-specific

discrimination provided by KLDA to be preserved, resulting in a richer and more discriminative embedding for downstream classifiers, such as SVM (with polynomial and RBF kernels) and RF. Through the combination of targeted feature selection and kernel fusion, improved accuracy, robustness, and computational efficiency were achieved for crop recommendations under varying environmental conditions. The combined feature set outperformed the baseline of raw data and individual embedding methods, with accuracy increases of 18.5% for Polynomial SVM, 10.1% for RBF SVM, and 4.7% for RF. The suggested kernel fusion framework serves as a high-performance solution for precision agriculture, facilitating dependable decision-making without the need for extensive field trials or laboratory experiments.

The paper is organized as follows: In the following "Materials and Methods" section, we detail the nutrient dataset comparing Maize and Barley, the process of feature selection, and the kernel-fusion pipeline, which includes RBF-KPCA, Nystroem-KLDA, and a seven-dimensional embedding. This section also covers the SVM models (both polynomial and RBF) and RF models. Then "Results and Discussion" highlights classification accuracies across different feature spaces and includes partial-dependence analyses. Finally, last section "Conclusions" provides a summary of the key findings and their implications for precision agriculture.

MATERIAL AND METHOD

Dataset Description and Preprocessing

This study used a publicly available crop recommendation dataset consisting of 3,867 instances, which includes a variety of soil characteristics and meteorological parameters collected across four seasons. A detailed list of all features is presented in Table 1. Each instance includes a combination of soil chemical properties—such as pH, nitrogen (N), phosphorus (P), potassium (K), sulfur (S), and zinc (Zn)—along with soil color and various seasonal meteorological variables. These meteorological factors include maximum and minimum temperature, humidity (QV2M), total precipitation (PRECTOTCORR), wind speed (WS2M), cloud cover (CLOUD_AMT), and surface pressure (PS), all recorded across four seasons: winter, spring, summer, and autumn. The target variable is a categorical label indicating the most suitable crop for cultivation under the specified conditions. In this study, we studied a total of 1006 data samples consisting of maize and barley samples.

Table 1. Description of Dataset Features

Category	Column Name	Description
Soil Properties	Soilcolor	Color of the soil (categorical)
	pH	Soil pH value
	K, P, N	Potassium, phosphorus, and nitrogen content
	Zn, S	Zinc and sulfur content
Seasonal Weather Data	QV2M-W/Sp/Su/Au	Specific humidity at 2 meters (Winter, Spring, Summer, Autumn)
	T2M_MAX-W/Sp/Su/Au	Maximum temperature (°C) for each season
	T2M_MIN-W/Sp/Su/Au	Minimum temperature (°C) for each season
	PRECTOTCORR-W/Sp/Su/Au	Precipitation amount (mm/day) for each season
Other Meteorological Data	WD10M	Wind direction at 10 meters
	GWETTOP	Surface soil moisture content
	CLOUD_AMT	Cloud coverage percentage
	WS2M_RANGE	Wind speed range at 2 meters
	PS	Surface pressure
Target Variable	label	Recommended crop label (e.g., Barley, Maize, Groundnut, etc.)

Feature Extraction

To effectively capture both the global nonlinear structure and the discriminative information from six scaled soil nutrient variables, we utilize two complementary kernel methods: KPCA and KLDA (Hekmatmanesh et al., 2020). The results from these methods are then combined to create a unified feature vector.

KPCA enhances classical PCA by mapping original input vectors $x_i \in \mathbb{R}^{\boldsymbol{d}}$ into a high-dimensional Hilbert space $\mathfrak{H}$ using a nonlinear feature map φ, explicitly computing φ. Instead, it defines the Gram matrix as follows (Equation 1):

$$K_{ij} = \langle \varphi(x_i), \varphi(x_j) \rangle = \exp\left(-\gamma \| x_i - x_j \|^2\right) \quad (1)$$

using an RBF kernel with width parameter γ. Centering in feature space yields (Equation 2)

$$\widetilde{K} = K - \mathbf{1}_n K - K\mathbf{1}_n + \mathbf{1}_n \boldsymbol{K} \mathbf{1}_n \quad (2)$$

where $\mathbf{1}_n$ is the nxn matrix with all entries $1/\,n$. Then it is solved the eigenvalue problem (Equation 3)

$$\widetilde{K}\mathrm{v} = \lambda \mathrm{v} \quad (3)$$

and retain the top $m = 6$ eigenvectors $\{v_1, \ldots, v_6\}$ corresponding to the largest eigenvalues $\lambda_1 \geq \cdots \geq \lambda_6$. The kth KPCA component for a new sample x is computed as (Equation 4)

$$[v_k]^T[K(x, x_1), \ldots, K(x, x_n)]^T \quad (4)$$

producing a six-dimensional embedding that preserves the directions of greatest variance in the RBF-kernel feature space.

KPCA is an unsupervised method, whereas KLDA is a supervised approach aimed at maximizing the scatter between classes in relation to the scatter within classes, but within a nonlinear kernel space. To make the computation manageable, we begin by approximating the RBF mapping using the Nystrom method. This involves randomly selecting $p = 50$ basis points $\{z_j\}$, compute the $n \times p$ submatrix K_{np} of kernel values $\kappa(x_i, z_j)$, and form a low-rank feature map (Equation 5)

$$\tilde{\varphi}(x) \approx K_{nz}{K_{zz}}^{-1/2} \quad (5)$$

where K_{zz} is the $p \times p$ kernel matrix among the basis points. On this approximated p-dimensional representation, we perform standard LDA: compute the class-conditional means μ_c and overall mean μ, then solve (Equation 6)

$$\boldsymbol{w} = arg\ max_w\ w^T S_B w / w^T S_W w \quad (6)$$

with between-class scatter $S_B = \sum_c N_c(\mu_c - \mu)(\mu_c - \mu)^T$ and within-class scatter $S_W = \sum_i (x_i - \mu_{yi})(x_i - \mu_{yi})^T$. The leading eigenvector $\boldsymbol{w}$ of $S_W^{-1} S_B$ defines a single discriminant axis; projecting each sample onto this axis yields the one-dimensional KLDA feature.

These two kernel projections together provide a seven-dimensional feature set—six KPCA components and one KLDA coordinate. This set captures both the dominant nonlinear variance and the supervised class separation essential for effective crop recommendation.

Feature Fusion

After extracting nonlinear embeddings using KPCA and identifying a discriminant axis through KLDA, we combined these complementary feature sets into a unified representation. Let (Equation 7),

$$z_i^{KPCA} \in \mathbb{R}^m \quad (7)$$

denote the m-dimensional KPCA embedding of sample i and let (Equation 8)

$$z_i^{KLDA} \in \mathbb{R} \quad (8)$$

its corresponding one-dimensional KLDA score. We then create a concatenated feature vector (Equation 9)

$$f_i = [z_i^{KPCA} \quad z_i^{KLDA}]^T \in \mathbb{R}^{m+1} \quad (9)$$

from these components. This fusion step retains the global variance structure captured by KPCA while also emphasizing the supervised class separation highlighted by KLDA (Guan et al., 2024). To ensure numerical stability and maintain a balanced influence of each component, we optionally apply an additional standardization step (Equation 10),

$$\tilde{f}_i = StandardScaler(f_i) \quad (10)$$

where each dimension of f_i is centered to zero mean and scaled to unit variance across the training set.

The resulting vectors in $(m + 1)$ dimensions serve as input to downstream classifiers, merging diverse nonlinear perspectives into a unified feature space that improves model robustness and predictive accuracy.

A SVM is a supervised learning algorithm that creates optimal hyperplanes to separate data points in high-dimensional space. It is especially effective when the number of dimensions exceeds the number of samples and is applicable for both linear and non-linear classification tasks.

Support Vector Machine

SVMs, introduced by Vapnik (1995), are widely used for classification and prediction in supervised machine learning. They work by identifying the optimal decision boundary, known as a hyperplane, which maximizes the margin between different classes. This hyperplane is defined in a high-dimensional space, and the goal is to make the distance (or margin) to the nearest data points, known as support vectors, as large as possible. The concepts of support vectors, the hyperplane, and the margin are illustrated in Figure 1.

SVM handles non-linearly separable data by introducing slack variables and a regularization parameter C, which trades off margin width against classification errors and permits some points to lie within or across the margin. If even this "soft-margin" formulation fails to separate the classes in the original input space, SVM employs the kernel trick—using polynomial, RBF, or other kernel functions—to implicitly project the data into a higher-dimensional feature space where a linear separation may exist. In its dual form, SVM training reduces to solving a convex quadratic program over Lagrange multipliers, guaranteeing a unique global optimum. At inference time, each new sample is classified by computing the sign of the weighted sum of its kernel similarities to the support vectors plus a learned bias term.

The margin of a linear SVM classifier is defined as the perpendicular distance between the two supporting hyperplanes that touch the nearest training points from each class. When the separating hyperplane is expressed as (Equation 11)

$$\mathbf{w}.\boldsymbol{x} + b = 0 \quad (11)$$

the margin width is given by (Equation 12)

$$d = 2/\|\mathbf{w}\| \quad (12)$$

where $\|\mathbf{w}\|$ denotes the Euclidean norm of weight vector **w.** The same Euclidean metric (Equation 13),

$$((\boldsymbol{x_2} - \boldsymbol{x_1})^2 + (\boldsymbol{y_2} - \boldsymbol{y_1})^2)^{1/2} \quad (13)$$

applies to any points (x_1, y_1) and (x_2, y_2) in $\mathbb{R}^2$, and is used both in computing the margin and in measuring inter-point distances in the input space.

A labeled dataset of N examples (Equation 14),

$$\{(x_i, y_i)\}_{i=1}^N \quad (14)$$

with feature vectors $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iq}) \in \mathbb{R}^q$ and class labels $y_i \in \{-1, +1\}$, is used for training. The parameters by solving a constrained optimization problem that maximizes this margin while enforcing correct (or soft-margin) classification of the training points.

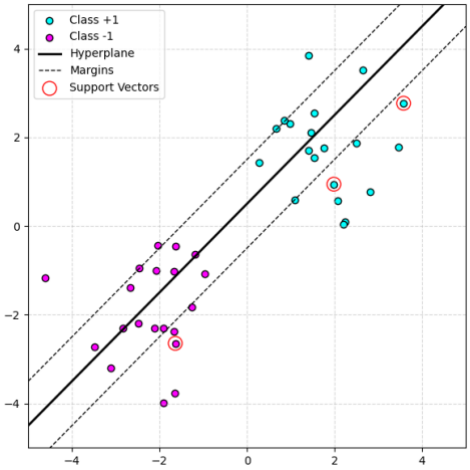

Figure 1. Illustration of SVM: Hyperplane, Margin and Support Vectors

SVMs were initially developed to solve linearly separable classification problems; however, such ideal scenarios rarely occur in real-world applications. To handle nonlinear class boundaries, SVMs use kernel functions that implicitly map the original feature vectors into richer, higher-dimensional spaces. In these spaces, it becomes easier to separate the classes with a linear hyperplane. During the training process, SVMs identify a subset of data points known as support vectors. These support vectors determine the position and orientation of the optimal decision boundary, with those having larger weights exerting a greater influence on the boundary's placement (Piccialli & Sciandrone, 2022).

In the soft-margin formulation, the hyperplane parameters are found by minimizing (Equation 15)

$$\boldsymbol{\Phi}(\mathbf{w}, \boldsymbol{\xi}) = 1/2 \, \|\mathbf{w}\|^2 + C \sum_{i=1}^{N} \xi_i \quad (15)$$

subject to (Equation 16)

$$y_i(\mathbf{w}.\mathbf{x}_i + b) \geq 1 - \boldsymbol{\xi}_i \, , \boldsymbol{\xi}_i \geq 0, \; i = 1, \ldots, N \quad (16)$$

where each $\mathbf{x}_i \in \mathbb{R}^n$ is a training feature vector, $y_i \in \{\mp 1\}$ is its class label, C controls the trade-off between margin width and training error, and ξ_i are the slack variables.

Kernel methods replace the inner product $\mathbf{x}_i.\mathbf{x}_j$ with a kernel function (Equation 17)

$$k(\mathbf{x}_i.\mathbf{x}_j) = \Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j) \quad (17)$$

where $\Phi(.)$ is the implicit feature map. The resulting decision function takes the form (Equation 18)

$$f(\mathbf{x}) = \sum_{j=1}^{N} \alpha_j y_j k(\mathbf{x}_j, \mathbf{x}) + b \quad (18)$$

in which the α_j are Lagrange multipliers, b is the bias term defining the hyperplane ofset, and the support vectors are those training points with nonzero α_j.

In this study, we employ two classic SVM kernel functions. First, the polynomial kernel of degree d is defined as (Equation 19)

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) \; = \; (\gamma \, \mathbf{x}_i^{\top} \mathbf{x}_j + C)^d \quad (19)$$

where γ controls the influence of higher-order terms, C is a constant offset, and d is the polynomial degree.

Second, we use the RBF kernel, given by (Equation 20)

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) \; = \; exp\left(-\gamma \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right) \quad (20)$$

where the parameter γ determines the "spread" of the Gaussian and thus the flexibility of the decision boundary.

Random Forest

RF is an ensemble algorithm based on bagging that constructs multiple decision trees using bootstrap samples from the training data. At each split in a tree, only a random subset of features is considered, which promotes diversity among the trees and reduces correlation between the individual learners. During inference, each tree casts a "vote" for the predicted class, and the final output of the forest is determined by majority voting. This process helps lower variance compared to using a single decision tree.

The model's hyperparameters—such as the number of trees, the maximum tree depth, and the number of features sampled at each split—control the trade-off between bias and variance. Feature importance can be measured by the mean decrease in impurity (MDI), which averages the reduction in Gini impurity (or entropy) contributed by each feature across all splits and trees. This combination of bootstrapping, random feature selection, and aggregation creates a robust classifier that generalizes well to unseen data and offers insights into feature relevance (Ibrahim 2022).

DATA ANALYSIS, EXPERIMENTAL STUDIES AND RESULTS

Data analysis

To gain a clearer understanding of the dataset's distributional properties and structure, exploratory data analysis (EDA) was conducted. The objective was to identify any skewness, outliers, and seasonal trends within both soil and weather-related features. As illustrated in Figure 2, several soil attributes—including phosphorus (P), potassium (K), and sulfur (S)—displayed strongly right-skewed distributions with significant outliers. This suggests that normalization or transformation of these data may be necessary before modeling. In contrast, features such as pH and nitrogen (N) exhibited more symmetrical or multimodal distributions, reflecting the underlying variability across different regions.

Figure 3 shows the correlation heatmap for the numerical features in the dataset. Strong positive correlations are observed among the seasonal temperature variables (e.g., T2M_MAX-W, T2M_MAX-Sp, T2M_MAX-Su), indicating that higher temperatures during different seasons are closely related. Additionally, specific humidity values (such as QV2M-W, QV2M-Sp, etc.) also exhibit strong mutual correlations.

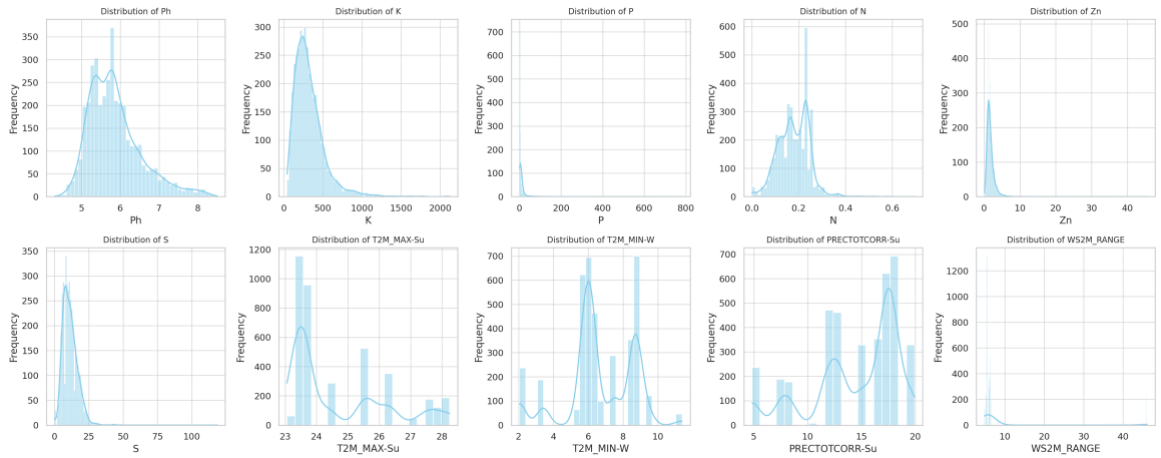

Figure 2. Kernel density estimates of the empirical distributions for the selected features used in feature selection

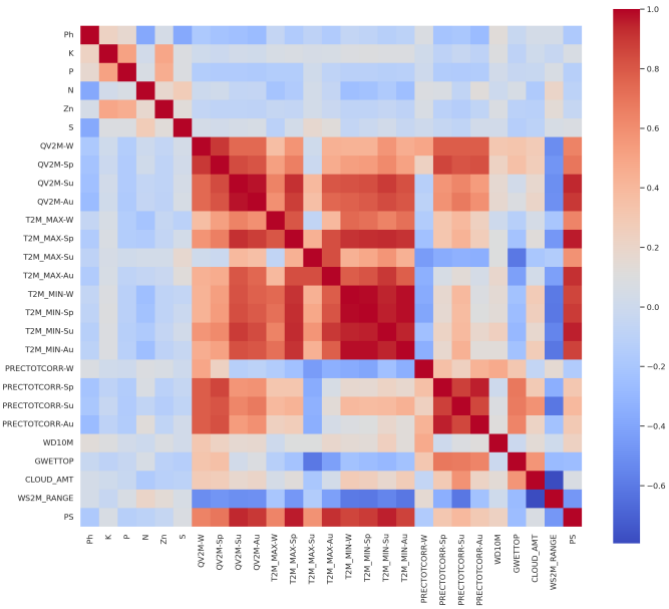

Figure 3. Pearson correlation heatmap of all numerical features considered for model development(red = positive, blue = negative.)

To highlight seasonal patterns, boxplots were created to display maximum temperatures and precipitation levels across the four seasons. Figure 4 illustrates that summer temperatures are significantly higher than those in winter and spring. Meanwhile, Figure 5 shows that precipitation peaks during the summer and decreases in winter, confirming the dataset's seasonal granularity.

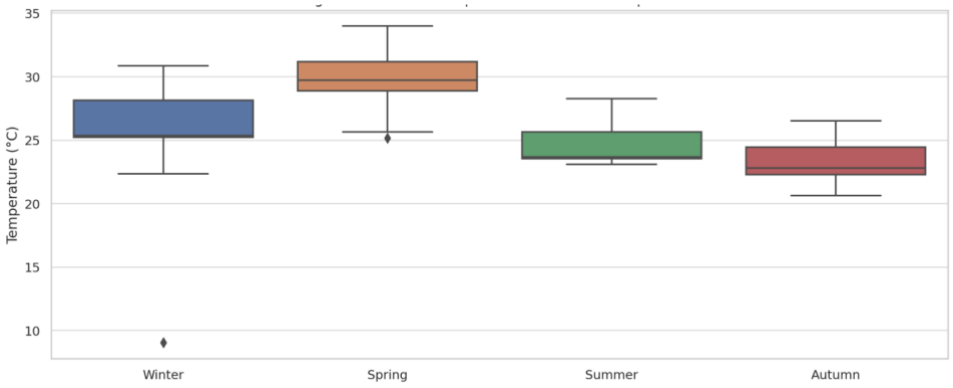

Figure 4. Seasonal distribution of daily maximum temperatures (°C) shown as boxplots for Winter, Spring, Summer, and Autumn.

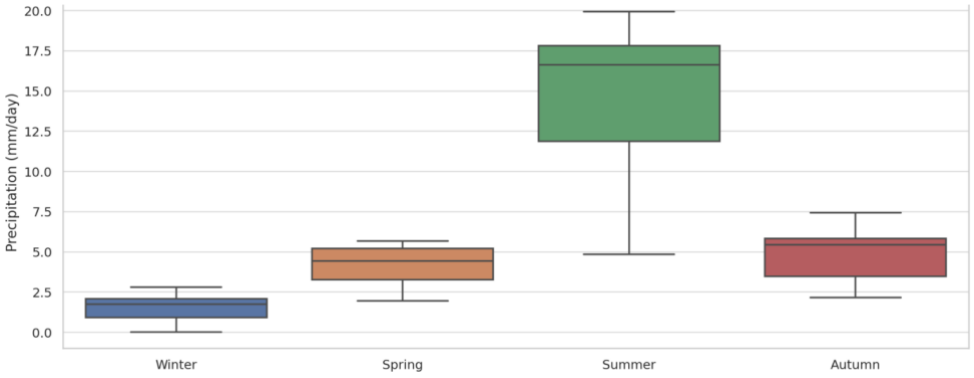

Figure 5. Boxplots of daily precipitation (mm/day) for Winter, Spring, Summer and Autumn.

Feature Importance and Feature Selection

First, we conducted feature importance analysis using a RF model, with results shown in Figure 6. Potassium (K) received the highest importance score, making it the single most influential feature for distinguishing Maize from Barley. Nitrogen (N) and zinc (Zn) hold significant discriminative power, closely followed by sulfur (S) and soil pH, which rank fourth and fifth, respectively. This highlights the importance of these chemical properties in crop recommendations. Phosphorus (P) has been an important predictor, though it is somewhat less influential than pH. In addition to the six core nutrients, certain soil color categories—specifically brown, black, and red—provide moderate additional information. In contrast, seasonal weather variables have been less significant in terms of their importance. These findings indicate that feature-selection efforts should focus on K, N, Zn, S, pH, and P. Additionally, soil color should be considered a secondary categorical feature while meteorological factors take on a supportive role.

The heatmap given in Figure 7 illustrates that the strongest interaction effects occur between potassium (K) and itself (highlighted in the K–K diagonal), as well as between potassium and phosphorus (K–P); these cells are the brightest in the matrix. Nitrogen's interaction with potassium (N–K) is also notable, indicated by a lighter green color. In contrast, zinc (Zn) shows the weakest interactions with all other nutrients (depicted in deep purple), signifying minimal synergistic effects. Additionally, sulfur (S) and pH both exhibit small but non-zero interactions with potassium and nitrogen (marked in mid-range blues), suggesting minor pairwise contributions. Conversely, the nearly empty columns for zinc, sulfur, and pH indicate that their mean SHAP interaction values with the other features are close to zero.

Experimental Studies

We retained the six most influential predictors from the feature importance plot (Figure 6) potassium (K), nitrogen (N), zinc (Zn), sulfur (S), soil pH, and phosphorus (P) discarding less important variables. We then created a reduced dataset with these predictors and the crop label, evaluating our crop recommendation framework.

In this framework (Figure 8), the raw nutrient data for Maize and Barley is loaded and cleaned. Following this, the data is split into stratified training and test sets. All six soil-nutrient features are then standardized before the feature extraction step is applied, followed by feature fusion. Finally, two different SVM kernels—polynomial and RBF and RF were trained on these embeddings.

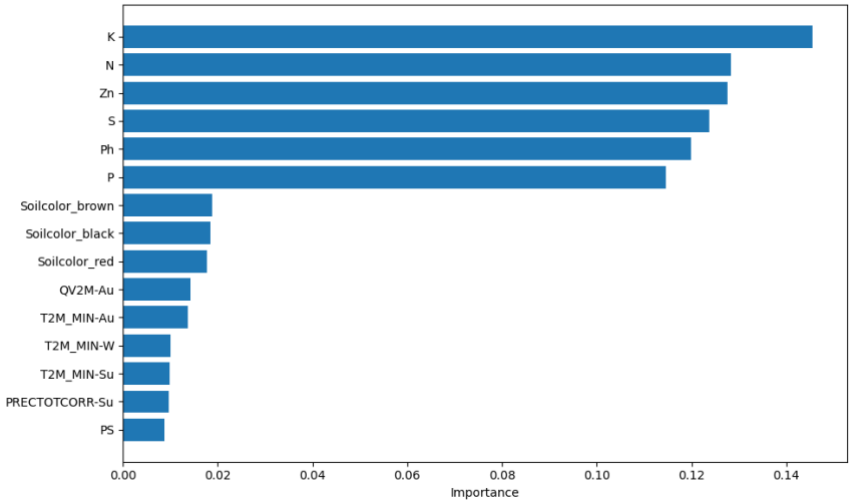

Figure 6. Top 15 feature importances as estimated by a Random Forest classifier (mean decrease in impurity).

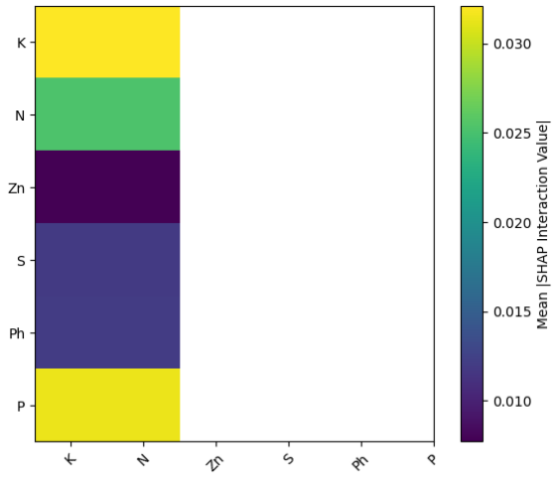

Figure 7. Mean absolute SHAP interaction values computed across the test set. Mean SHAP interactions (brighter = stronger): K and P show the largest interactions, while Zn and several auxiliary features are weak.

Explanation of tese experiments were given in Table 2. In each experiment, the six soil nutrient variables (K, N, Zn, S, pH, and P) were standardized to have a mean of zero and a variance of one. The dataset was then divided into an 80/20 stratified train/test split. Model performance was evaluated using test set accuracy.

```
A[Load & Clean Data] --> B[Train/Test Split]
B --> C[Scale Features]
C --> D[Feature Extraction]
D --> E[Concatenate- Feature Fusion]
E --> F[SVM Classifiers]
F--> H[Evaluate: Accuracy, F1-Score, Confusion Matrix]

subgraph SVM_Kernels [ ]
   direction LR
   K1[Linear]
   K2[Polynomial]
   K3[RBF]
   K4[Sigmoid]
end
I --> SVM_Kernels
```

Figure 8. Pipeline flowchart of the proposed crop-recommendation system (data cleaning → scaling → kernel feature extraction → fusion → SVM classification → evaluation).

RESULTS AND DISCUSSION

In the original feature space (Figure 9 Top left), Barley and Maize samples show significant overlap when plotted along any two raw nutrient axes. This suggests that the six-dimensional soil nutrient variables alone do not offer a clear distinction for classifying the crops. Although there are slight clustering tendencies—such as Maize samples having slightly higher potassium values—no decision boundary appears in this raw space.

In contrast, the KPCA embedding (Figure 9 Top right) unfolds the data into a six-dimensional nonlinear manifold. Two principal components reveal some structural separation: Maize points tend to cluster on one side of the KPCA1–KPCA2 plane, while Barley points gather on the opposite side. However, there is still substantial overlap between the classes, indicating that unsupervised variance maximization alone only partially clarifies the underlying class boundaries.

The KLDA projection (Figure 9 Bottom left) presents a markedly different view: by extracting a single supervised discriminant axis, the samples of Barley and Maize are nearly perfectly aligned along this one-dimensional line. Low KLDA scores are predominantly associated with Barley, while high scores correspond primarily to Maize. This outcome confirms that supervised, kernel-based class separation can effectively concentrate essential discriminant information into a single coordinate.

Table 2. Explanation of four experimental studies

Experiment	Description
Raw Features with Four SVM Kernels	Train four SVMs directly on the scaled six-dimensional nutrient data without feature extraction. Serves as reference performance.
KPCA (m=6) with Four SVM Kernels	Apply RBF-kernel PCA to the scaled nutrients, retaining **m=6** components that capture the bulk of nonlinear variance. Train four SVM kernels on this 6-D KPCA embedding to assess unsupervised variance reduction benefits.
KLDA (1D) with Four SVM Kernels	Approximate the RBF feature map with **p=50** Nystroem basis functions, then perform LDA to extract a single supervised discriminant axis (KLDA). Four SVMs are trained on this 1-D KLDA output to measure the impact of supervised kernel projection.
KPCA (6D) + KLDA (1D) Fusion with Four SVM Kernels	Concatenate the six KPCA components with the one KLDA score into a 7-D fused feature vector. Train four SVM variants on this hybrid embedding to evaluate whether combining unsupervised and supervised kernels yields additional gains over KPCA or KLDA alone.

The fused KPCA+KLDA space (Figure 9 Bottom right) integrates six components from KPCA with one score from KLDA, resulting in a seven-dimensional embedding. In this space, the first two KPCA axes, along with the KLDA axis, create a 3D scatter plot in which the class clusters are clearly separated. Maize and barley occupy distinct areas within this hybrid space, utilizing both global variance patterns and supervised discriminative power. This combination produces a representation that significantly enhances the accuracy of SVM classification compared to using either method alone.

As shown in Figure 10, using only raw nutrient features results in modest baseline accuracies: 0.599 for the Polynomial SVM, 0.663 for the RBF SVM, and 0.688 for the RF model. When applying KPCA alone, all three models experience a consistent boost in accuracy to 0.710. This demonstrates that unsupervised variance maximization leads to a reliable, though limited, improvement in performance. On the other hand, applying KLDA alone results in significant improvements for the two SVM variants: the Polynomial SVM increases by 11.9% (from 0.599 to 0.670), while the RBF SVM improves by 7.1% (from 0.663 to 0.710). In contrast, RF experiences a slight decline of 1.2% (from 0.688 to 0.680), indicating that the single supervised axis may discard some variance that tree-based methods find beneficial.

The feature fusion of KPCA with KLDA yields significantly better results across all models. The accuracy of the Polynomial SVM increases by 18.5%, rising from 0.599 to 0.710. The RBF SVM shows an improvement of 10.1%, increasing from 0.663 to 0.730, while RF's accuracy increases by 4.7%, going from 0.688 to 0.720. These consistent improvements demonstrate that merging unsupervised kernel variance with a supervised discriminant axis creates a more effective and discriminative feature space. This is particularly advantageous for SVMs using polynomial and RBF kernels in crop recommendation tasks.

To evaluate the impact of kernel-based feature extraction methods effectively, we chose a dataset that clearly displays significant nonlinear characteristics, as shown in our data scatter plots. This inherent nonlinearity helps to explain why the overall prediction accuracies given with the bar plot in Figure 10 tend to remain moderate on average.

Figure 11 illustrates the decision boundaries learned by polynomial and RBF SVMs on the fused KPCA+KLDA embedding, comparing KPCA Component 1 with Component 2, and coloring by KLDA score. the RBF SVM in the right panel produces a smoother, more localized boundary that tightly wraps around the high-density cluster in the center while correctly excluding many of the peripheral Barley points. The color gradient of the KLDA scores further illustrates how the RBF kernel utilizes supervised separability along the vertical axis to refine its decision boundaries in the KPCA plane. Overall, these plots confirm that the RBF SVM is more effective at leveraging the combined unsupervised and supervised feature space, leading to more accurate class separation compared to the polynomial kernel.

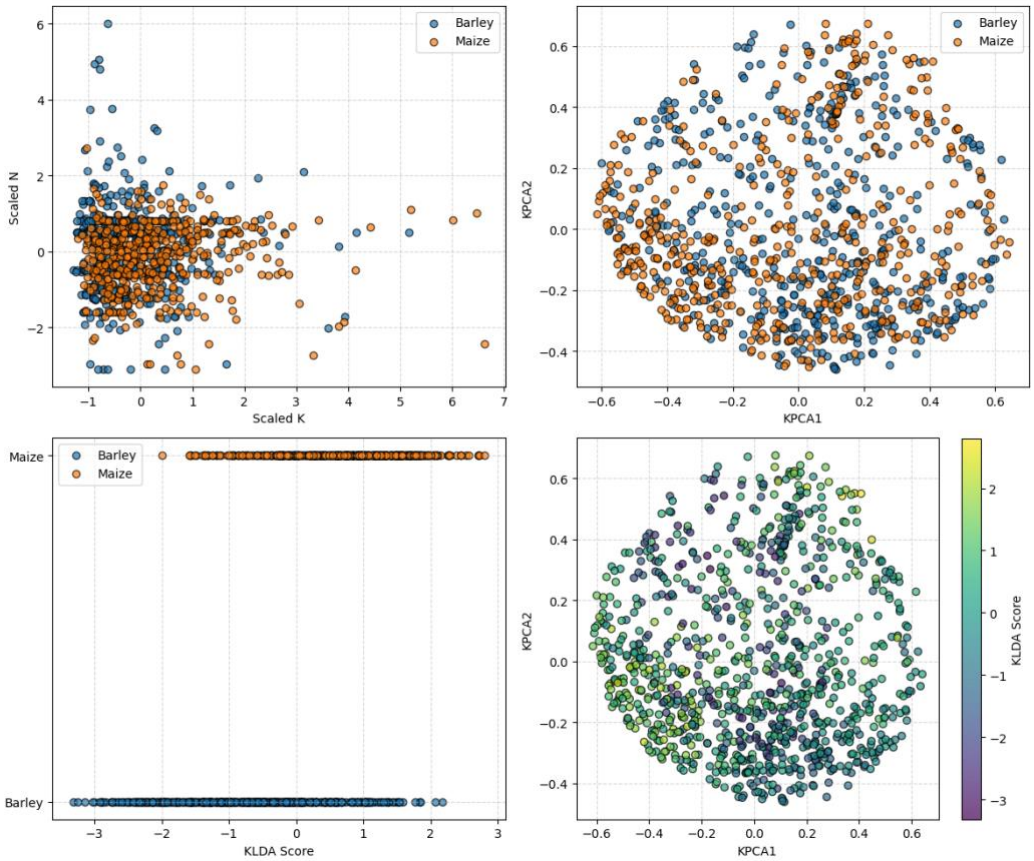

Figure 9. Data scattering plots. The top left shows raw K vs. N., the top-right plot shows the first two KPCA components, the bottom left displays the one-dimensional KLDA score and the bottom-right fuses KPCA1,2 colored by KLDA.

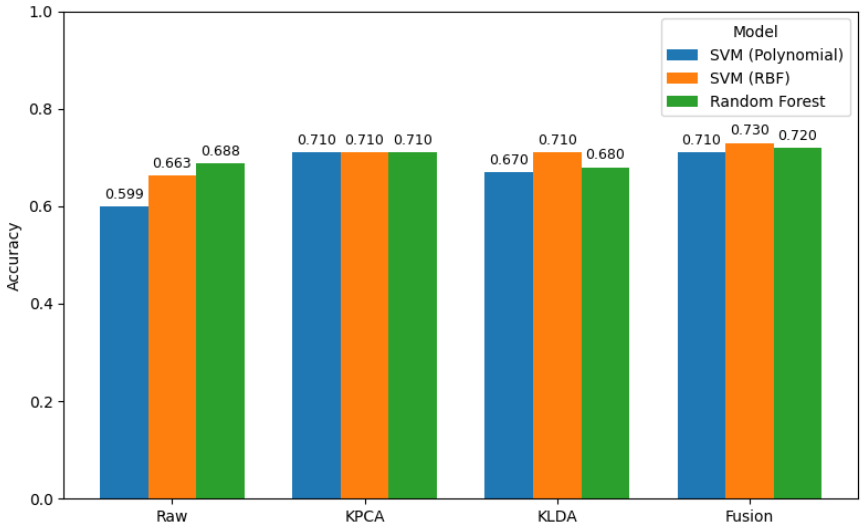

Figure 10. Accuracies of Polynomial SVM, RBF SVM, and RF using Raw, KPCA, KLDA, and fused KPCA+KLDA feature spaces.

The Partial Dependency Plots (PDPs) shown in Figure 12 illustrate that integrating the unsupervised KPCA components with the supervised KLDA axis creates a more complex and multidimensional decision surface in the feature fused space.The Polynomial SVM utilizes this fused space by balancing the negative influence of KPCA1 with the separation provided by KLDA. Meanwhile, the RBF SVM takes advantage of both the global variance captured by KPCA and the local discrimination achieved through KLDA to create more flexible decision boundaries. Additionally, the tree-based ensemble benefits from the increased variance represented by KPCA, leading to smoother, upward-trending partial dependence profiles compared to the case with only KLDA.

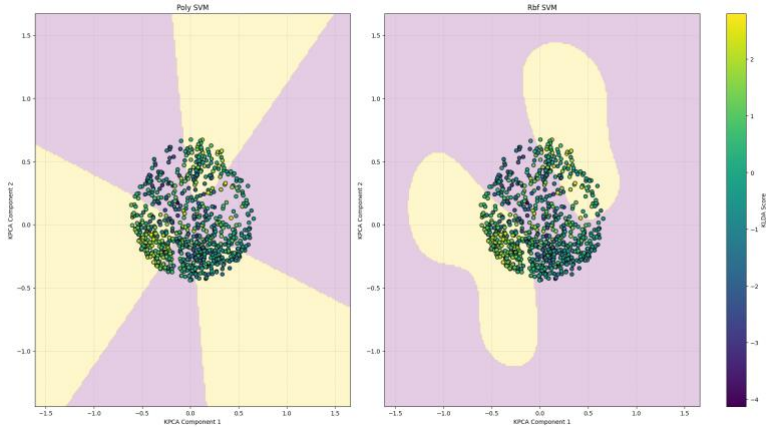

Figure 11. Decision boundaries of polynomial (left) and RBF (right) SVMs on the fused KPCA + KLDA feature space (points colored by KLDA score).

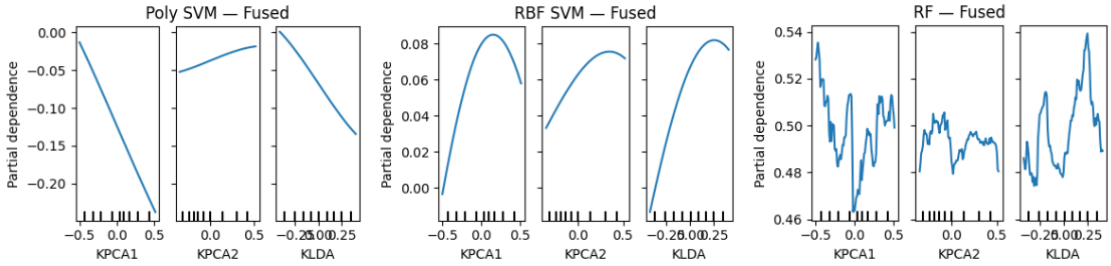

Figure 12. Pdp analysis on feature fused space for Poly SVM, RBF SVM and RF

CONCLUSION

Through our comprehensive experimentation, we have found that understanding the underlying data distribution is crucial when choosing a classification method. The raw nutrient scatter (Figure 9, top-left) shows that the six original features have significant overlap, which makes linear or low-complexity models unsuitable without further feature transformation. On the other hand, the KLDA projection (Figure 9, bottom-left) illustrates that even a single, well-selected discriminant axis can achieve nearly perfect class separation. However, relying solely on this supervised perspective overlooks important variance, as indicated by the modest decline in RF performance. Our results emphasize the transformative power of feature fusion. By combining unsupervised KPCA components with the supervised KLDA axis, we create a hybrid embedding that captures both the global nonlinear structure of the data and the specific boundaries between classes. This fused representation consistently outperforms each standalone approach, increasing Polynomial SVM accuracy by 18.5%, RBF SVM by 10.1%, and RF by 4.7% compared to raw features. It also yields the most flexible decision surfaces, particularly for RBF kernels (Figure 11) and richer partial-dependence profiles (Figure 12). Notably, the novel integration of KPCA and KLDA in this study represents, to our knowledge, the first demonstration of supervised-unsupervised kernel fusion applied to crop recommendation. This fusion maximizes the discriminatory information available to downstream classifiers while addressing the limitations of purely unsupervised or purely supervised transformations. In the context of recommending maize versus barley based on soil nutrients, our kernel feature fusion framework stands out as a powerful and generalizable approach for precision agriculture. It enables more accurate and robust decision-support models without the need for extensive field trials or manual feature engineering.

Compliance with Ethical Standards

Peer Review

This article has been reviewed by independent experts in the field using a rigorous double-blind peer review process.

Conflict of Interest

The authors declare no conflicts of interest.

Author Contributions

All authors contributed equally to the study design, data collection, analysis, and manuscript preparation.

Ethics Committee Approval

Ethical approval was not required for this study.

Consent to Participate / Publish

Written informed consent was obtained from all participants.

Funding

The authors declare that this study received no financial support.

Data Availability

The datasets generated during and/or analyzed in the current study are available from the corresponding author upon reasonable request.

Generative AI Statement

No generative AI tools were used in the writing, editing, data analysis, or figure preparation of this manuscript.

REFERENCES

Anowar, F., Sadaoui, S., & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne). *Computer Science Review*, *40*, 100378.

Bandara, P., Weerasooriya, T., Ruchirawya, T., Nanayakkara, W., Dimantha, M., & Pabasara, M. (2020). Crop recommendation system. *International Journal of Computer Applications*, *975*, 8887.

Barburiceanu, S., Meza, S., Orza, B., Malutan, R., & Terebes, R. (2021). Convolutional neural networks for texture feature extraction. Applications to leaf disease classification in precision agriculture. *IEEE Access*, *9*, 160085-160103.

Baudat, G., & Anouar, F. (2000). Generalized discriminant analysis using a kernel approach. *Neural computation*, *12*(10), 2385-2404.

Briscik, M., Dillies, M. A., & Déjean, S. (2023). Improvement of variables interpretability in kernel PCA. *BMC bioinformatics*, *24*(1), 282.

Boppudi, S., & Jayachandran, S. (2024). Improved feature ranking fusion process with Hybrid model for crop yield prediction. *Biomedical Signal Processing and Control*, *93*, 106121.

Cheng, X. (2024). A comprehensive study of feature selection techniques in machine learning models. *Available at SSRN 5154947*.

Çağlar, E. (2024). The impact of sectors on agriculture based on artificial intelligence data: a case study on G7 countries and Turkiye. *International Journal of Agriculture Environment and Food Sciences*, *8*(3), 486-494.

Getahun, S., Kefale, H., & Gelaye, Y. (2024). Application of precision agriculture technologies for sustainable crop production and environmental sustainability: A systematic review. The Scientific World Journal, 2024(1), 2126734.

Gosai, D., Raval, C., Nayak, R., Jayswal, H., & Patel, A. (2021). Crop recommendation system using machine learning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *7*(3), 558-569.

Guan, H., Ren, Y., Tang, H., & Xiang, J. (2024). Intelligent fault diagnosis methods for hydraulic components based on information fusion: review and prospects. *Measurement Science and Technology*, *35*(8), 082001.

Gunawan, M. I., Sitopu, J. W., Sechan, G., & Gunawan, I. (2024). Optimizing Crop Selection: A Multi-Criteria Decision Support System for Sustainable Agriculture. *International Journal of Enterprise Modelling*, *18*(3), 113-123.

Gür, Y. E. (2024). Innovation in the dairy industry: forecasting cow cheese production with machine learning and deep learning models. *International Journal of Agriculture Environment and Food Sciences*, *8*(2), 327-346.

Hekmatmanesh, A., Wu, H., Jamaloo, F., Li, M., & Handroos, H. (2020). A combination of CSP-based method with soft margin SVM classifier and generalized RBF kernel for imagery-based brain computer interface applications. *Multimedia Tools and Applications*, *79*(25), 17521-17549.

Huang, X., Wang, H., & Li, X. (2024). A multi-scale semantic feature fusion method for remote sensing crop classification. *Computers and Electronics in Agriculture*, *224*, 109185.

Ibrahim, S. A. (2022). Improving land use/cover classification accuracy from random forest feature importance selection based on synergistic use of sentinel data and digital elevation model in agriculturally dominated landscape. *Agriculture*, *13*(1), 98.

Jiao, L., Xie, C., Chen, P., Du, J., Li, R., & Zhang, J. (2022). Adaptive feature fusion pyramid network for multi-classes agricultural pest detection. *Computers and electronics in agriculture*, *195*, 106827.

Kahya, E., & Aslan, Y. (2024). Detection of artichoke on seedling based on YOLOV5 model. *International Journal of Agriculture Environment and Food Sciences*, *8*(1), 186-201.

Kaushal, S., Nayi, P., Rahadian, D., & Chen, H. H. (2022). Applications of electronic nose coupled with statistical and intelligent pattern recognition techniques for monitoring tea quality: A review. *Agriculture*, *12*(9), 1359.

Kempfert, K. C., Wang, Y., Chen, C., & Wong, S. W. (2020). A comparison study on nonlinear dimension reduction methods with kernel variations: Visualization, optimization and classification. *Intelligent Data Analysis*, *24*(2), 267-290.

Kusuma, C. G., Dharumarajan, S., Vasundhara, R., Gomez, C., Manjunatha, M. H., & Hegde, R. (2025). Predicting Soil Nutrient Classes Using Vis–NIR Spectroscopy to Support Sustainable Farming Decisions. *Land Degradation & Development*.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, *18*(8), 2674.

Linaza, M. T., Posada, J., Bund, J., Eisert, P., Quartulli, M., Döllner, J., ... & Lucat, L. (2021). Data-driven artificial intelligence applications for sustainable precision agriculture. *Agronomy*, *11*(6), 1227.

Mahesh, T. R., Thakur, A., Velmurugan, A. K., Khan, S. B., Gadekallu, T. R., Alzahrani, S., & Alojail, M. (2024). AgriFusion: A Low-Carbon Sustainable Computing Approach for Precision Agriculture Through Probabilistic Ensemble Crop Recommendation. *Computational Intelligence*, *40*(6), e70006.

Mishra, H., & Mishra, D. (2024). AI for data-driven decision-making in smart agriculture: From field to farm management. In *Artificial Intelligence Techniques in Smart Agriculture* (pp. 173-193). Singapore: Springer Nature Singapore.

Peng, P., & Zhao, Y. P. (2023). Robust semi-supervised discriminant embedding method with soft label in kernel space. *Neural Computing and Applications*, *35*(11), 8601-8623.

Piccialli, V., & Sciandrone, M. (2022). Nonlinear optimization and support vector machines. *Annals of Operations Research*, *314*(1), 15-47.

Ruano-Ordás, D. (2024). Machine learning-based feature extraction and selection. *Applied Sciences*, *14*(15), 6567.

Swaminathan, B., Palani, S., & Vairavasundaram, S. (2023). Feature fusion based deep neural collaborative filtering model for fertilizer prediction. *Expert Systems with Applications*, *216*, 119441.

Taye, M. M. (2023). Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. *Computers*, *12*(5), 91.

Upadhyay, N., Sharma, D. K., & Bhargava, A. (2025). 3SW-Net: A Feature Fusion Network for Semantic Weed Detection in Precision Agriculture. *Food Analytical Methods*, 1-17.

Vapnik, V.N. (1995). The Nature of Statistical Learning Theory. Springer, New York.

Yanardağ, A. B. (2025). Soil chemistry in apricot cultivation: evaluation of C and N dynamics by PCA and correlation analysis. *International Journal of Agriculture Environment and Food Sciences*, *9*(2), 360-373.

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends*, *1*(1), 56-70.