MJFAS
Malaysian Journal of Fundamental and Applied Sciences

Jamshed *et al.* | Malaysian Journal of Fundamental and Applied Sciences, Vol. 21 (2025) 2945–2957

RESEARCH ARTICLE

Unsupervised Learning for Crop Suitability Clustering Based on Soil Nutrients

Huma Jamshed[a], Urooj Waheed[a], Yusra Mansoor[a], Ahmad Hussain[b*]

[a]Department of Computer Science, DHA Suffa University, Ph-VII، DG-78, Off Khayaban-e-Tufail، Ext, Phase 7 Ext Karachi, 75500, Pakistan; [b]Faculty of Engineering & Applied Sciences, DHA Suffa University, Ph-VII، DG-78, Off Khayaban-e-Tufail، Ext, Phase 7 Ext Karachi, 75500, Pakistan

Abstract Soil-based crop recommendation plays a critical role in precision agriculture, especially under increasing climate uncertainty and resource limitations. This study proposes a clustering-based framework that leverages unsupervised machine learning to group crops according to soil parameters. A labelled dataset of 2,201 soil samples covering 22 crop types was analysed to uncover patterns linking soil profiles with crop suitability. The results reveal clear distinctions among crop groups, with generalist crops like rice and maize appearing across multiple clusters, while crops such as apple and grape form tighter, more specific groupings. These insights highlight natural affinities between soil chemistry and crop behaviour, offering a practical, data-driven basis for region-specific crop planning and soil resource optimization. The study contributes toward scalable, interpretable decision tools for sustainable agriculture, particularly in environments where efficient land and input management are critical. Unlike prior studies that employ clustering generically, this work comparatively evaluates multiple unsupervised algorithms under agricultural conditions, integrating soil nutrient dynamics into interpretable cluster formation.

Keywords: Clustering Analysis, Data mining, K-Means clustering, Crops segmentation, Precision agriculture, Soil crop compatibility.

Introduction

Background and Context

Agriculture continues to be the most important sector of the global economy [1]. It provides us with one of the most basic needs we have, food. Unfortunately, the agriculture sector remains highly volatile and is easily influenced by external factors, such as environmental and economic risks, including climate change, soil degradation, and biodiversity loss, which directly impacts commodity prices [2][3]. The increases in global population and urbanization also place significant stress on sustainable agriculture.

In order to address food demand, we need to understand what crops, environmental conditions and market conditions are happening in real time. To respond to this continuous agriculture these sustainable agricultural smart technologies needed to be integrated use in off-farm applications using IoT along with on-farm applications using Machine learning (ML). [4] The automation of farm wide action and being able to make decisions based visited performances data will help farmer's farm maximize both yields and profitability.[5].

Clustering analysis is an unsupervised ML process being used to identify trends and segmentation by attributes including crop types, soil types, and environmental data [6]. It permits classification of the optimum conditions for crop production, yield forecasting, and efficient resource allocation [7]. Early crop mapping, especially before harvest, has been shown to improve productivity by contributing to growth monitoring and yield forecasting, and by identifying areas of high productivity [8].

Clustering can allow segmentation of fields by soil type, moisture, and climate as well as group crops when similar growing conditions can create efficiencies in harvesting schedules and advance prediction of future trends. [9] When applied, clustering can provide stakeholders with the means for improved decisions to enhance crop management and total productivity in agriculture. [10]

***For correspondence:**
rosdiyana@uitm.edu.my

Received: 5 August 2025
Accepted: 21 Oct. 2025

Related Work

Internet of Things (IoT) and ML are continuously transforming the conventional agriculture practices to sustainable, data drive approach. The advancement of technology in this sector is able to address several issues such as soil degradation, climate variability, and water scarcity problem. For crop segmentation clustering analysis is increasingly being utilized, making it a valuable tool for optimizing resource allocation and predicting yields.

[11] integrates IoT and ML system for real time soil nutrient monitoring and crop recommendation. [12] presented a comprehensive review of IoT and ML based precision agriculture frameworks focused on predictive analytics for decision making. [13] developed a crop recommendation system that combines soil, weather, and grain storage data to suggest suitable crops. By analysing nutrient levels and climatic variables, the system helps farmers make data-driven decisions to improve yield and fertilizer efficiency.

[14] proposed a semi-supervised constrained K-Means clustering method to map soil texture in hilly areas using limited labelled samples along with topographical and land use features. [15] delineated management zones in olive groves using unsupervised methods like K-Means, hierarchical clustering, and DBSCAN, showing that the choice of clustering technique significantly affects the quality of zone delineation. [16] used ISO cluster unsupervised classification with Sentinel-2 and Landsat-8 data to predict soil nutrient indices such as phosphorus, iron, and pH, achieving recognition rates of 97%, 94.05%, and 69% respectively, thereby supporting early fertility assessment. [17] proposed Deep Crop Clustering (DCC), a deep unsupervised clustering method that uses contractive learning and nearest–farthest neighbor sorting to map crops without labelled data, achieving better performance than conventional methods. [18] used explainable AI (XAI) to analyse how soil chemical and microbial factors affect microbial respiration's temperature sensitivity (Q_{10}), revealing that microbial communities play a dominant role under climate stress.

[19] developed a hybrid model that integrates ensemble learning (bagging, boosting, stacking) with K-Means clustering to recommend optimal crops for varying environmental conditions, improving predictive robustness. [20] examined a range of ML classifiers within a crop recommendation implementation based on clusters, where crops were grouped using K-Means clustering, and classifiers recommended 2–3 crops from clusters to evaluate each system's accuracy without increasing complexity. [21] developed an ML approach to evaluate soil fertility and moisture content, allowing for tailored irrigation and fertilization applications to evaluate the precision of farming practices.

[22] developed a framework for dynamic zone delineation that combines NDVI, elevation, and soil texture data that utilizes clustering and a geographically weighted regression to account for spatial yield variability [23] in unsupervised ML applications for irrigation management of rice fields with the use of spatial PCA and multispectral data. Indicators of seasonal soil moisture variability were identified through two management zones where significant differences could be documented throughout multiple cycles. [24] created a clustering approach to incorporate climate and NDVI time series data with SPEI, designating seasons as dry, normal, or wet prior to clustering with a K-Means algorithm. This clustering approach provided more correspondence between soil properties and soil zones than traditional clustering methods even in dry years.

[25] created a hybrid spectral–cLHS approach to sample soil organic carbon. Their hybrid method increases covering the environmental axes while providing improved quality training data by combining spectral clustering with conditioned Latin hypercube sampling for SOC prediction. [26] investigated sunflower yield differences due to crop year and planting density across a productivity zone. The data indicated positive responses of sunflower yields to density optimizations in high and mid-zone productivity levels during wet years, while low-zone productivity levels had poor responses. The work indicated that factors could enter into adaptive field management with weather patterns changing.

Materials and Methods

This research leverages unsupervised ML algorithms to cluster crops based on soil parameters. The objective is to identify sets of crops that require similar soil nutrients for optimal growth, thereby enabling better agricultural planning. The dataset for this research was sourced from the Kaggle agricultural database. The research methodology follows the Cross-Industry Standard Process for Data Mining (CRISP-DM), which is internationally recognized framework for all stages of the data mining process, including data collection, data preparation, model building, and assessment as illustrated in Figure 1.

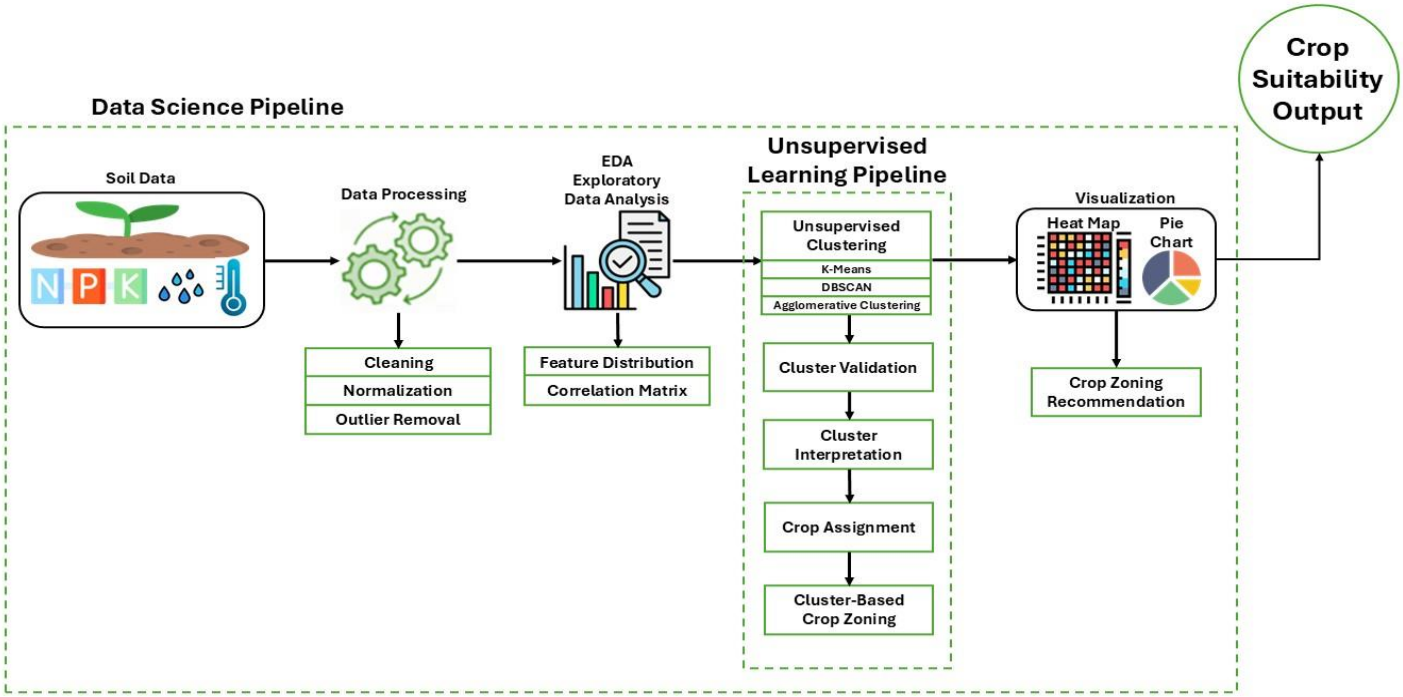

Figure 1. Architecture of proposed Methodology

As shown in Figure 1, the collected dataset is pre-processed to handle missing values, noise, and outliers. Then feature engineering was performed to select the relevant variable and reduce the dimensionality of the dataset. The dataset is then split into training, validation, and testing subset using random split technique. For the model selection, three different unsupervised algorithms, K-means clustering, DBSCAN, and Agglomerative Clustering, were applied to identify the best algorithm that could recognize the inherent pattern of the data. For K-means clustering, the optimal number of clusters was determined using the Elbow Method, Silhouette Analysis, and Gap Statistic Method. Together, these methods ensure both the appropriate cluster count and robust internal cohesion for the soil crop dataset.

Data Collection and Preprocessing

The free and open-sourced dataset was taken from Kaggle in CSV format. It contains data for about 22 different crops, including both fruits and vegetables such as rice, maize, pigeon peas, kidney beans, chickpeas, and moth beans. Each crop type was then assigned a unique class label, which is represented mathematically in Equation 1.

$$C = \{c1, c2, c3, c4, \ldots, cn\} \tag{1}$$

The collected dataset *D* is then processed to enhance its quality and make it suitable for the ML algorithm. The pre-processing step involves identifying and replacing incomplete, inaccurate, irrelevant, or noisy data. To handle missing values, row wise deletion was applied. This approach removes records with missing values to maintain the integrity of the remaining dataset. The dataset *D* is represented as a matrix of size *m x n* where *m* is the number of rows and *n* is the number of columns. Let x_{ij} represent the value of the j^{th} feature in the i^{th} row. Let $R \subseteq \{1,2,\ldots,m\}$ be the set of rows containing missing values. The dataset *D'* after removing rows with missing values is given by equation 2.

$$D' = \{x_i \in D \,/\, i \in! R\} \tag{2}$$

The outliers are removed by utilizing a technique called the Z-scores. Outliers distort the results of machine learning models and reduce their accuracy. The Z-score for a data point f in an attribute *A'* is calculated using the mean μ and standard deviation σ of *A*, as expressed in equation 3.

$$Z = \frac{A' - \mu}{\sigma} \tag{3}$$

The final stage of data preprocessing involves the encoding for categorical variable C can be represented mathematically by equation 4.

$$f: C \rightarrow Z \quad (4)$$

Exploratory Data Analysis (EDA)

A comprehensive Exploratory Data Analysis (EDA) was done on pre-processed datasets in order to understand structure, relationship and hidden patterns. The process started with statistical data summary including mean, median and standard deviations in order to explore the data distribution. Table 1 presents the statistical summary of the different feature sets available in the given data set.

Table 1. Statistical Summary of crop dataset

Statistic	Nitrogen	Phosphorus	Potassium	Temperature	Humidity	pH	Rainfall
Count	2200.00	2200.00	2200.00	2200.00	2200.00	2200.00	2200.00
Mean	50.55	53.36	48.14	25.61	71.48	6.46	103.46
Std	36.91	32.98	50.64	5.06	22.26	0.77	54.95
Min	0.00	5.00	5.00	8.82	14.25	3.50	20.21
25%	21.00	28.00	20.00	22.76	60.26	5.97	64.55
50%	37.00	51.00	32.00	25.59	80.47	6.42	94.86
75%	84.25	68.00	49.00	28.56	89.94	6.92	124.26
Max	140.00	145.00	205.00	43.67	99.98	9.93	298.56

For further visual interpretations histogram were generated understand the distribution and variability of each feature in the dataset. The histograms for pH, temperature, and humidity were bell-shaped, reflecting a normal distribution. Therefore, these features are well-suited for use in ML models without requiring any transformation. The histogram of potassium and phosphorus showed skewness and demands transformation. The humidity histogram shows two main peaks, representing two different groups in the dataset. These groups could represent crops grown in different weather conditions or regions. The distribution of crop features is shown in figure 2.

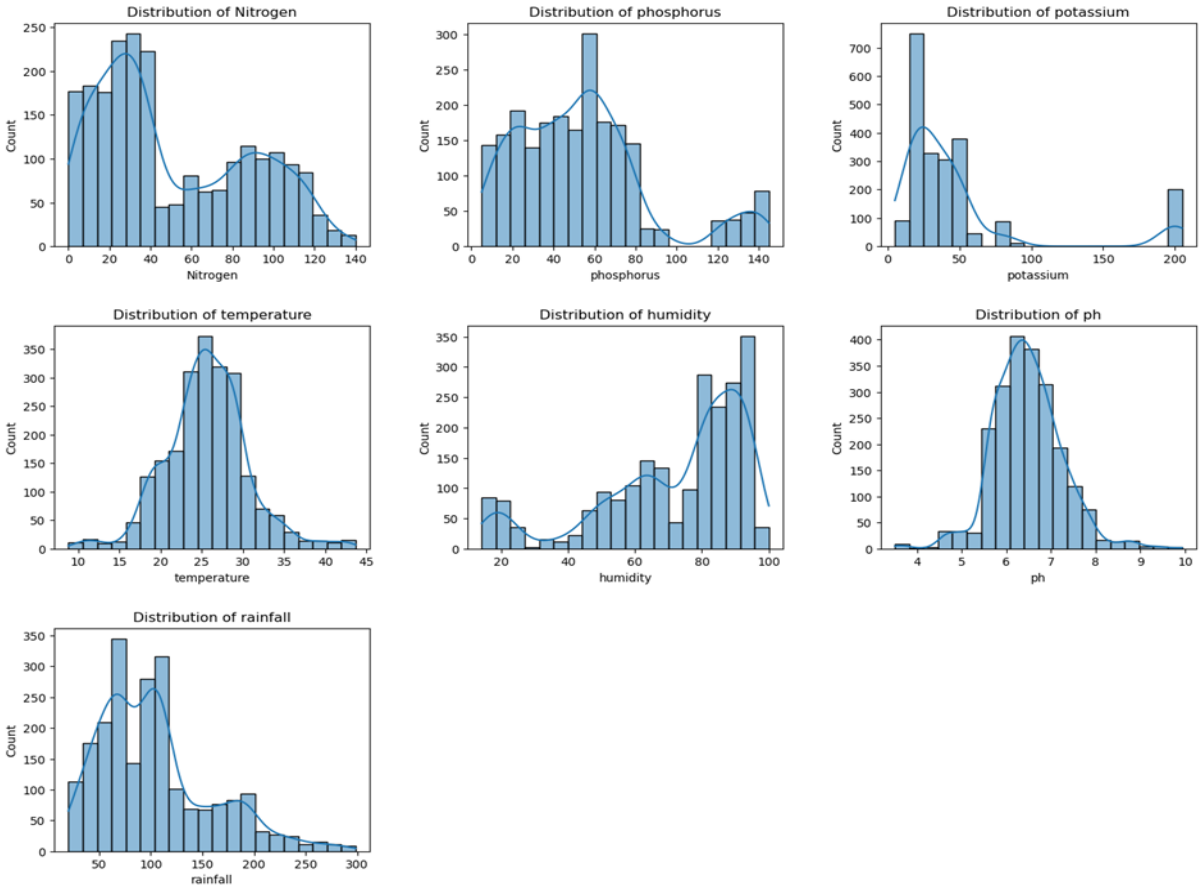

Figure 2. Histogram for Crop Feature Distribution

To identify relationships among the features in the pre-processed dataset, a correlation matrix was generated to determine the final feature set that would be used to train the ML model. Figure 3 presents correlation matrix heatmap for the given data set. Each cell of the matrix represents the Pearson correlation coefficient between two features. The colour of the heatmap indicate the strength and direction of relation of two features. The analysis reveals phosphorus and potassium have a strong correlation indicating a fundamental dependency for instance soil composition. Independent relationship exists among other feature set, which can be beneficial in predictive modeling as they may provide unique information. The statistical and correlation insights, particularly highlighted the relevance of soil nutrients and pH values.

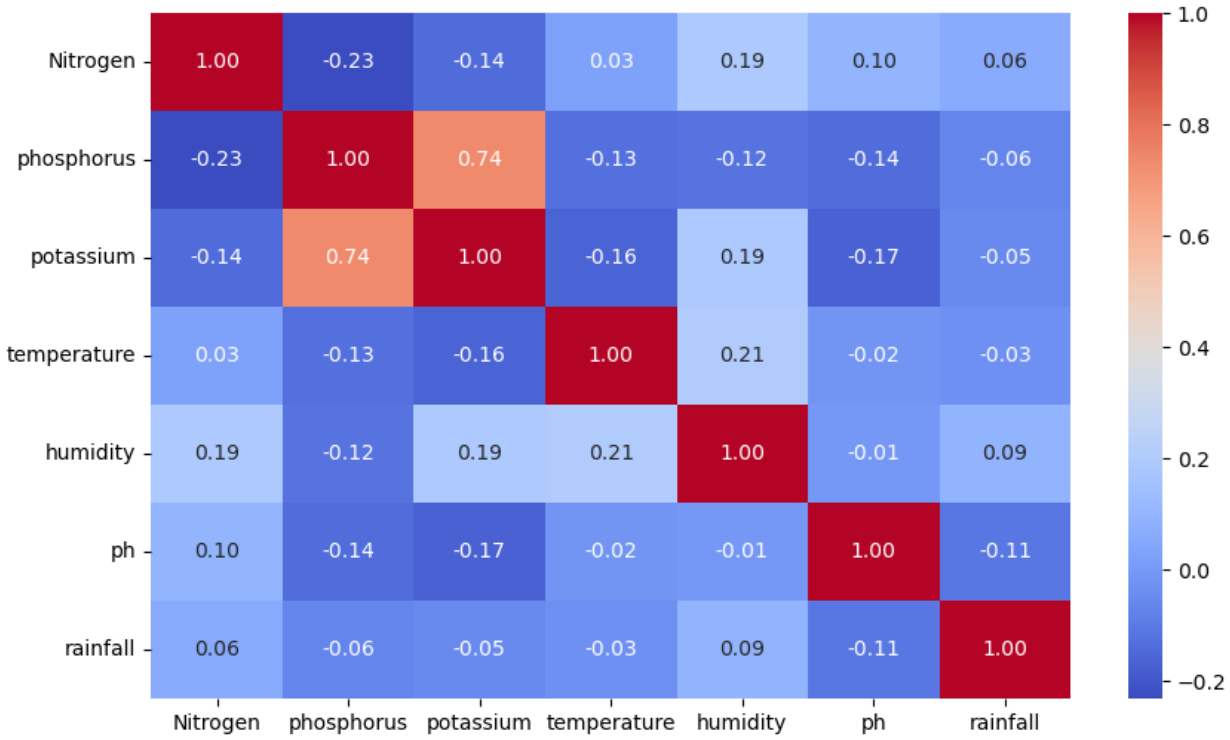

Figure 3. Correlation matrix of soil features

Feature Engineering

Feature engineering allows to selected feature sets that not only show meaningful variation across samples but also reflected significant properties that influence crop growth. Feature involved extracting relevant features expected to significantly influence model outcomes. This process enhances the dataset by creating new features from existing ones and enriching the data structure. Mathematically, let $F=\{f_1,f_2,\ldots,f_n\}$ represent the set of original features in the dataset. Through feature engineering, a new feature set F 'is generated as represented by equation 5

$$F'=\{g(f_1, f_2, \ldots, f_k)\} \quad (5)$$

Where:

g is a feature transformation function

Data Split Training and Testing Sets

In this phase the data is split into three distinct subsets for ML algorithms i.e., the training set (D_{train}), validation set (D_{val}), and test set (D_{test}), ensuring that there is no overlap between them. The split is typically based on proportions such as 70-80% for training, 10-15% for validation, and 10-15% for testing, satisfying the condition as shown in equation 6.

$$D_{train} + D_{val} + D_{test} = 1 \quad (6)$$

The training set is used to train the model by minimizing the loss function and learning the parameters, and the model is also trained with the validation set for hyperparameter tuning and model performance monitoring, as indicated in Equation (7). The test set is used for the final evaluation of the model's performance on new data.

$$D_{train}, D_{val}, D_{test} = randomSplit(D, [training, Validation, testing], seed) \quad (7)$$

Model Selection and Configuration

With the processed dataset, we applied three unsupervised learning algorithms:K-Means, DBSCAN, and Agglomerative Clustering are used. This selection is based on algorithms capability to address explicit data characteristics and clustering requirements.

K-Means is a partitioning clustering method that minimizes intra-cluster variance. It is extremely efficient for datasets with numerical features and offers clear, non-overlapping clusters K-Means minimizes the sum of squared distances (SSD) between data points and their respective cluster centroids. Mathematically, this is expressed in equation 8.

$$J = \sum_{i=1}^{k} \sum_{j \in Ci} ||xi - \mu i||^2 \quad (8)$$

J is the total SSD for k clusters and C_i is the set of points assigned to cluster i. μ_i is the centroid of cluster. The optimal number of clusters is determined using techniques such as the Elbow Method, Silhouette Analysis, and Gap Statistics.

DBSCAN works well for datasets that have clusters in unusual shapes and can identify noise and outliers, and also does not rely on the user to define the number of clusters. DBSCAN generates groups by grouping points that are close to each other, and consider those in regions with lower densities as noise. The key parameters to set for DBSCAN are epsilon point ϵ, which specifies the maximum allowable distance between two data points for them to be treated as part of the same neighborhood, along with the minimum count of neighboring points needed for an area to qualify as a dense cluster. Any point that has at least the specified minimum number of neighbors within this distance ϵ is regarded as a core point can be defined as a core point. Clusters can grow by repeatedly including density-connected points, as can be seen in equation 9.

$$Core\ Point: |N(p)| \geq MinPts \quad (9)$$

Agglomerative Clustering is a hierarchical approach where every data point is treated as its own cluster at the beginning, and then clusters are progressively combined based on a measure of similarity. It outputs a dendrogram to show how the clusters relate to each other in a hierarchical fashion. The linkage criterion gives a distance between clusters. The common metrics are Single Linkage, which has a distance of the closest points of two clusters, Complete Linkage which has a distance of the furthest points of two clusters, and Average Linkage which has the average distance between all of the points in two clusters. The criteria for merging is stated in equation 10 of article. This diminishes within cluster variance and maximizes between cluster separation.

$$D(A, B) = min_{a \epsilon A, b \epsilon B} ||a - b|| \quad (10)$$

Results and Discussion

The clustering models were applied to a dataset of 2201 soil samples classified for 22 different crops. Key attributes include NPK (Nitrogen, Phosphorus, Potassium) levels, temperature, humidity, pH, and rainfall. The primary objective is to cluster crops that can thrive on similar soils based on these features. The simulation environment used to cluster crops was implemented in Anaconda Jupyter Notebook. For simulation purpose only the features containing numerical values of the soil's chemical properties were included. Nitrogen, Phosphorus, Potassium and pH were therefore selected as four-dimensional feature space. Although the dataset also included environmental variables such as temperature, humidity, and rainfall, but these were excluded from the feature sets to preserve emphasis on basic soil chemistry. The selected four features were the most influential indicators of soil fertility and directly determine crop suitability, while climatic parameters vary seasonally and may obscure underlying nutrient-based relationships. Figure 3 further justified the selection of feature set by showing weak associations between the environmental factors and the soil nutrients. The outcomes of the chosen clustering algorithms are discussed as follow;

K-means Clustering

In this study we utilized K-means as the first choice for crop clustering. For accurate clustering results in K-mean analysis number of clusters to be formed should be known in advance. After detailed literature review, we applied three techniques for determine the optimal number of clusters for K-means clustering including Elbow method, Silhouett method and Gap Statistics method.

Elbow method evaluates total intra-cluster variation or the within cluster sum of squares WSS as a function of the number of clusters. The average Silhouette method, calculates the average silhouette score for different numbers of clusters (k), with the optimal k being the one that maximizes the silhouette score.

The Gap Statistics method compares intra-cluster variation for different k values against expected values from a null reference distribution, where the optimal number of clusters is the one that maximizes the gap.

The results of the Elbow, Silhouette, and Gap Statistics methods are presented in Figures 4, 5, and 6, respectively.

Elbow method as shown in figure 4 gives four clusters for the given dataset. Likewise, silhouette method as shown in figure 5 gives four clusters. Gap Statistics method also suggested a maximum of four clusters. Based on these findings, four clusters were set for K-means algorithm analysis.

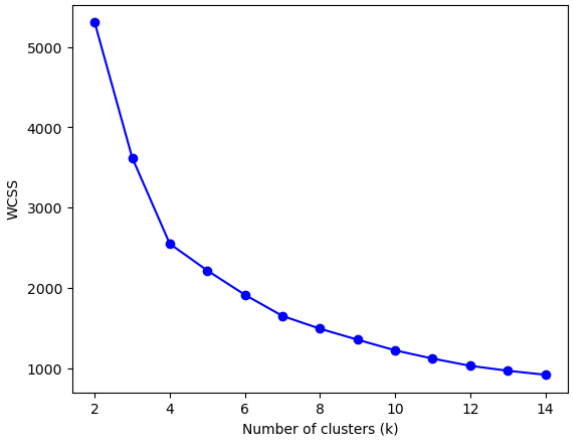

Figure 4. Optimal number of clusters using Elbow Method

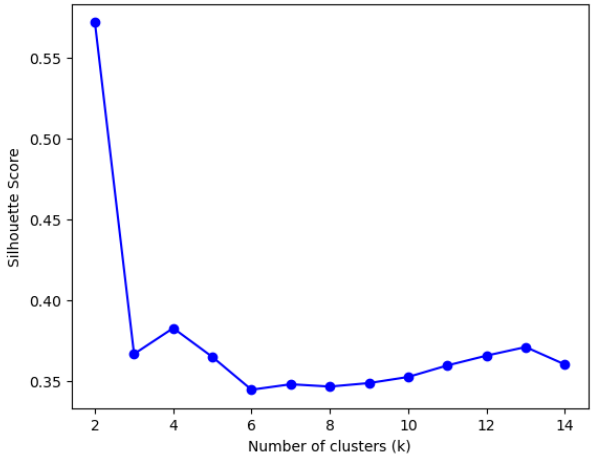

Figure 5. Optimal number of clusters using Silhouette Method

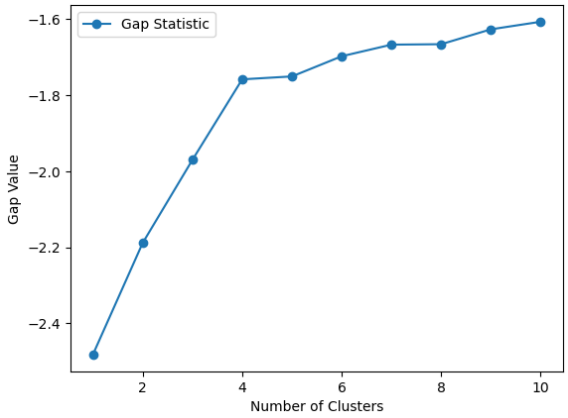

Figure 6. Optimal number of clusters using Gap Method

The analysis of K-means clustering on selected feature sets of the soil dataset using four clusters is shown in figure 7. The algorithm groups crop with similar soil characteristics into a single cluster. Some crops for example rice, maize, chickpea, and black gram appeared in multiple clusters because the data set contains varieties of rice with different soil preferences, while others were found in only one cluster due to more specific soil requirements. The presence of these crops in multiple clusters suggests that the clusters are not mutually exclusive and that the features used for clustering is overlapping. Similarly fruits like grapes and apple are grouped exclusively into a single cluster, indicating that these crops have distinct environmental and nutrient requirements. These results show how the clustering analysis groups crops that can grow in many soil types along with those crops which requires specific soil conditions.

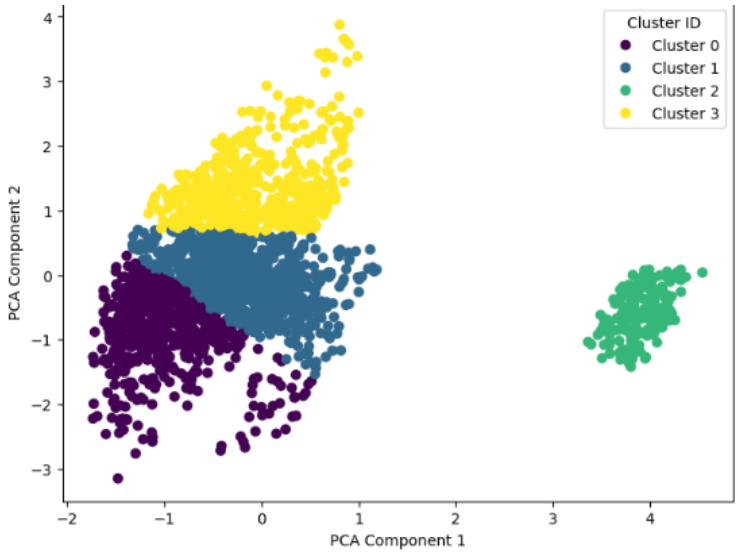

Figure 7. K-Mean Clustering of crop data into four groups

"Cluster0" in K-means, represents group of crops (rice, maize, chickpea, mung bean, pigeon peas, lentil, and jute) which requires soil with high nitrogen, moderate phosphorus and potassium, and a neutral pH. This cluster is ideal for leguminous and fiber crops. "Cluster1" in K-means, represents crop group (coffee, maize, chickpea, and pigeon peas) which require soils with balanced nutrient levels and slightly acidic pH and these crops grow well with moderate rainfall and temperature. "Cluster2" categorized crops which requires soil having low nitrogen but very high phosphorus and potassium levels with a slightly acidic pH. Such soil is well suited for fruit bearing crops like grapes and apples. Finally, the last cluster "Cluster3" contains crops (rice, maize, coconut, lentil, and mango) having acidic soils and lower nutrient concentrations. The distinct ranges of nitrogen (N), phosphorus (P), potassium (K), and pH values across four clusters is shown in Figure 8, each representing different soil conditions suitable for various crops as represented in Table 2.

Table 2. Crop suitability mapping based on K-Means soil cluster characteristics.

Cluster	N_min	N_max	P_min	P_max	K_min	K_max	ph_min	ph_max
Cluster0	58	140	5	95	15	84	5.01	7.99
Cluster1	0	78	5	80	5	85	4.19	6.70
Cluster2	0	78	5	80	5	85	6.28	8.76
Cluster3	0	40	120	145	195	200	5.51	6.49

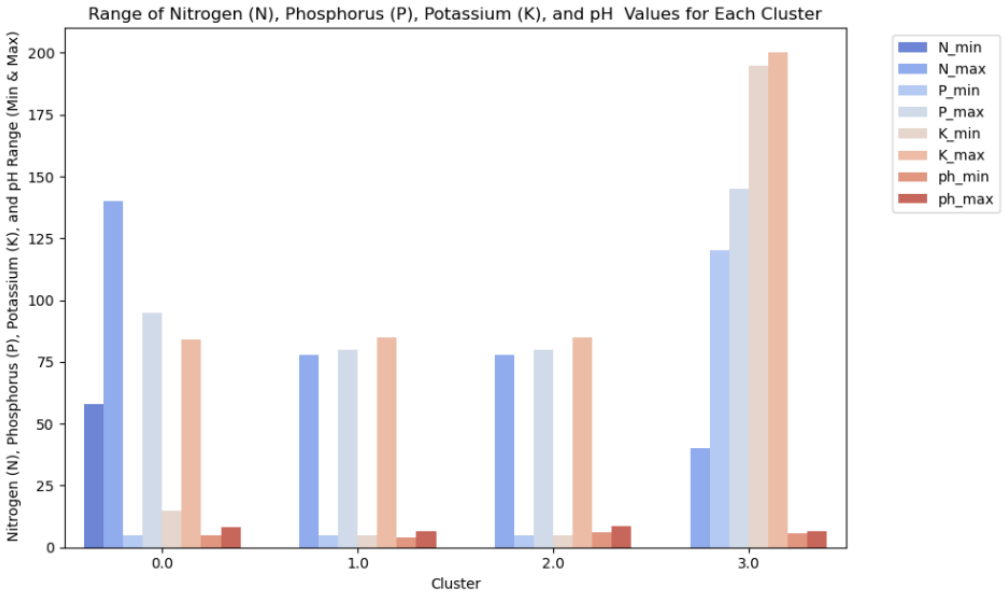

Figure 8. Cluster-wise distribution of N, P, K, and pH values derived from K-Means analysis

Determining which crops are in the same cluster can be useful in customizing agricultural practices and interventions. If one of the clusters is linked to a soil type or climate, for example, the crops within that cluster, will have similar needs or adapt to climate and/or soil alike allowing for more precise farming practices. The pie chart (Figure 9) displays the pattern of 20 different crops drawn from four clusters giving a visual representation of how different crops grouped within each of those clusters. Each slice of the pie chart is the proportion within a particular cluster, and the size of the slice represents the number of crops that are in that cluster.

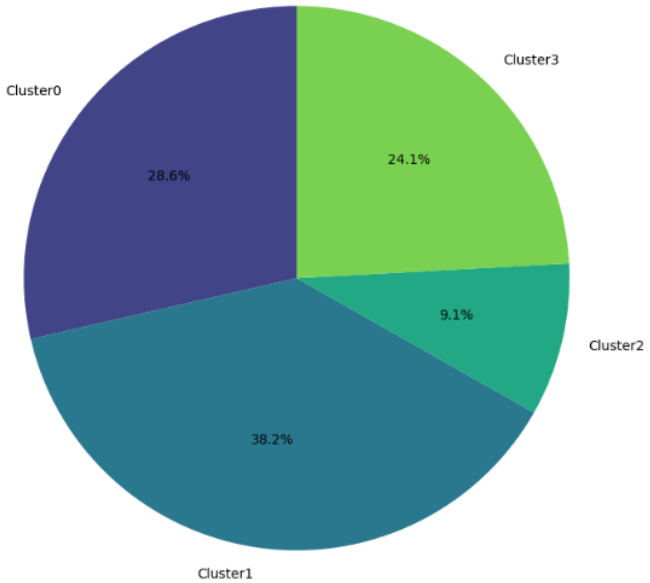

Figure 9. Distribution of 20 crops across four clusters based on soil suitability

DBSCAN Clustering

The same dataset was utilized for DBSCAN clustering and was used to validate the results obtained from K-means. DBSCAN method excels at identifying noise or outlier points, making it particularly suitable for real-world datasets where irregularities and anomalies are common. DBSCAN's ability to detect clusters of arbitrary shapes is also beneficial, especially when clusters are not spherical or evenly distributed.

Four groups were formed by the DBSCAN algorithm namely -1, 0, 1, and 2 as shown in figure 10. Data points having few neighboring points or were too distant from other points were marked as outliers and was grouped under cluster -1. Whereas clusters 0, 1, and 2 represent the three distinct clusters identified within the dataset. These clusters group crops with similar features based on soil characteristics chosen during training phase. Points within each cluster are closer to one another according to the epsilon distance parameter, forming dense regions.

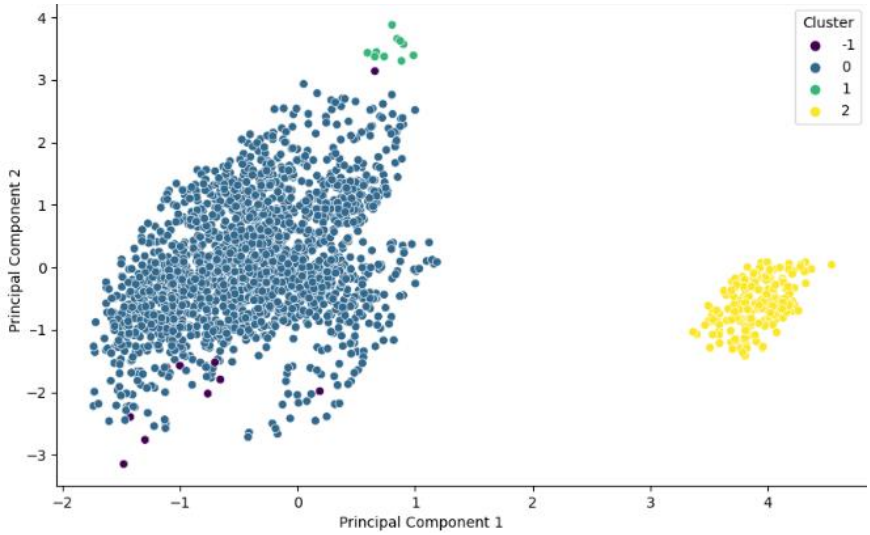

Figure 10. DBSCAN clustering results showing three dense clusters and one outlier group (Cluster -1) based on soil features

In DBSCAN cluster 0 is the largest and most diverse, encompassing a wide variety of crops. Cluster 1 uniquely classifies moth beans, which is due to a small but distinct set of features. Cluster 2 includes grapes and apple similar to K-means results, highlighting their unique soil requirements compared to other crops. Also, crops like orange, mango, and pomegranate are also the part of clusters 2. This consistency suggests that these crops share well defined soil features that make them easier to group.

Agglomerative clustering

Agglomerative clustering is a hierarchical clustering technique that further assesses the dataset as compared to the other clustering algorithms. It treats each data point as its own cluster and successively merges the closest clusters until a stopping criterion is met. This method provides a tree-like structure known as dendrogram that reveals relationships between clusters.

In the proposed study this technique marks Cluster 0 as the most diverse cluster, containing a wide range of crop like rice, maize, chickpea, kidney beans, pigeon peas, moth beans, mung bean, black gram, lentil, pomegranate, mango, orange, papaya, coconut, jute. Whereas the Cluster 1 includes legumes (such as kidney beans, moth beans, and lentils) and fruit-bearing crops (including pomegranate, mango, orange, and coconut). Cluster 2 represents a mix of grains, fruits, and cash crops. Figure11 represents the hierarchical clustering of the crop dataset.

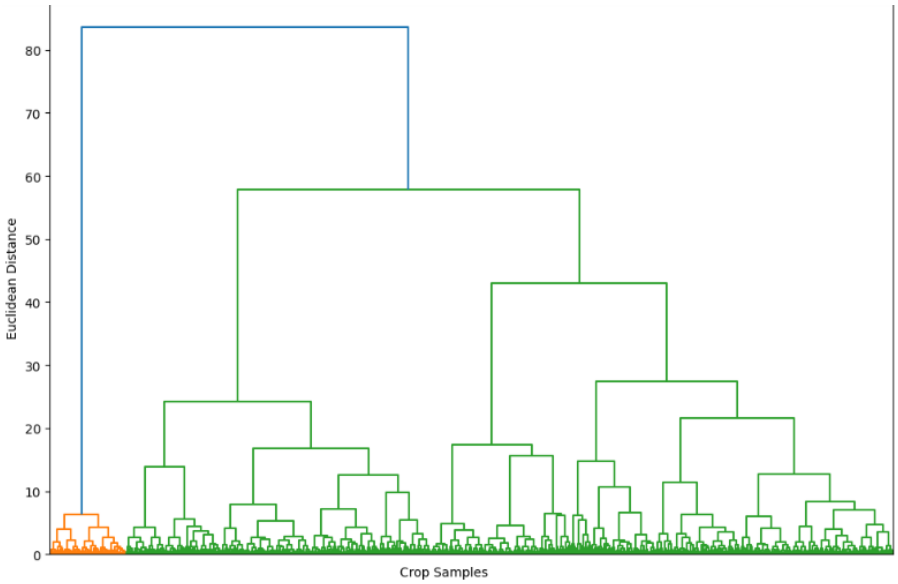

Figure 11. Hierarchical clustering of the crop dataset

The utilization of three different clustering algorithms helped in cross validation of the clustering outcomes, offering more robust conclusions about the grouping of crops in the dataset. All three algorithms identified similar groupings for many crops, with noticeable overlaps in clusters.

Agglomerative clustering results were closely aligned with clustering results of K-means. The cluster 0 of these two algorithm results had a broad range of crops partially similar to DBSCAN cluster 0. Whereas the cluster 1 in all three algorithms particularly focuses on legumes and fruits. Cluster 2 in K-means and Agglomerative clustering is comparable to DBSCAN's Cluster 2, representing that certain crop like apples and grapes are steadily grouped together.

Other than the visual interpretations for optimal number of clusters, we used Silhouette Coefficient measure to check how each crop fits in its cluster. The average silhouette scores for K-means, DBSCAN and Agglomerative clustering are 0.56, 0.49 0.52 respectively. This analysis confirmed that most crops were appropriately grouped within their respective clusters. The slight overlap observed for rice and maize appearing in multiple clusters reflects agronomic adaptability rather than algorithmic limitation.

Conclusions

In this investigation, we applied three unsupervised clustering algorithms, K-Means, DBSCAN, and Agglomerative clustering to assess the relationships between soil properties and crop suitability. The results identified numerous clusters where statistically significant soil nutrient properties, pH ranges, and crop types are closely associated with each other.

Through the comparative analysis of the three clustering algorithms, our research uncovered interpretable consistent crop groups based on feature set selected. The proposed data driven framework can assist agronomist with farmland planning, especially in resource-limited regions facing climate variability.

In our future work, we will focus on enhancing the result accuracy and depth of clustering results through the integration of additional feature sets such as rainfall patterns, irrigation data, and remote sensing visuals. In addition, we will investigate the incorporation of explainable AI within the proposed framework to offer greater transparency and trust in the decision-making process, or for other forms of stakeholder engagement. Overall, the proposed framework provides a scalable and interpretable platform for precision agriculture, feeding into the development of sustainable and resilient food systems within changing environments.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgment

This research was funded by DHA Suffa University, Karachi, Pakistan. The authors wish to express their gratitude to the DHA Suffa University for its technical and financial support.

References

[1] Singh, K., Yadav, M., Barak, D., Bansal, S., & Moreira, F. (2025). Machine-learning-based frameworks for reliable and sustainable crop forecasting. *Sustainability, 17*(10), 4711.
[2] Senoo, E. E. K., Anggraini, L., Kumi, J. A., Karolina, L. B., Akansah, E., Sulyman, H. A., Mendonça, I., & Aritsugi, M. (2024). IoT solutions with artificial intelligence technologies for precision agriculture: Definitions, applications, challenges, and opportunities. *Electronics, 13*(10), 1894.
[3] Afzal, H., Amjad, M., Raza, A., Munir, K., Villar, S. G., Lopez, L. A. D., & Ashraf, I. (2025). Incorporating soil information with machine learning for crop recommendation to improve agricultural output. *Scientific Reports, 15*(1), 8560.
[4] Eze, V. H. U., Eze, E. C., Alaneme, G. U., Bubu, P. E., Nnadi, E. O. E., & Okon, M. B. (2025). Integrating IoT sensors and machine learning for sustainable precision agroecology: Enhancing crop resilience and resource efficiency through data-driven strategies, challenges, and future prospects. *Discover Agriculture, 3*(1), 1–34.
[5] Yang, H., Lim, H., Moon, H., Li, Q., Nam, S., Kim, J., & Choi, H. T. (2022). Simple optimal sampling algorithm to strengthen digital soil mapping using the spatial distribution of machine learning predictive uncertainty: A case study for field capacity prediction. *Land, 11*(11), 2098.
[6] Folorunso, O., Ojo, O., Busari, M., Adebayo, M., Adejumobi, J., Folorunso, D., Ayo, F., Alabi, O., & Olabanjo, O. (2025). GeaGrow: A mobile tool for soil nutrient prediction and fertilizer optimization using artificial neural networks. *Frontiers in Sustainable Food Systems, 9*, 1533423.
[7] Ennaji, O., Vergütz, L., & El Allali, A. (2023). Machine learning in nutrient management: A review. *Artificial Intelligence in Agriculture, 9*, 1–11.
[8] Khan, M. H. (2022). *Modelling the long-term impact of modernized irrigation systems on soil water and salt balances, and crop water productivity*. Doctoral dissertation, Massey University.
[9] Pham, T.-H., Acharya, P., Bachina, S., Osterloh, K., & Nguyen, K.-D. (2024). Deep-learning framework for optimal selection of soil sampling sites. *Computers and Electronics in Agriculture, 217*, 108650.
[10] Salloum, S. A., Masa'deh, R., Al-Zoghb, A. M., & Shaalan, K. (2025). Optimizing crop recommendation using machine learning: An analytical study based on smart agricultural data. In *Generative AI in Creative Industries* (pp. 637–650). Springer Nature Switzerland.
[11] Islam, M. R., Oliullah, K., Kabir, M. M., Alom, M., & Mridha, M. F. (2023). Machine learning enabled IoT system for soil nutrients monitoring and crop recommendation. *Journal of Agriculture and Food Research, 14*, 100880.
[12] Saki, M., Keshavarz, R., Franklin, D., Abolhasan, M., Lipman, J., & Shariati, N. (2024). Precision soil quality analysis using transformer-based data fusion strategies: A systematic review. *arXiv*, arXiv:2410.18353.
[13] Sethi, S., & Lakhina, U. (2024). Intelligent crop selection and soil nutrient management using machine learning. In *2024 International Conference on Computational Intelligence and Computing Applications (ICCICA)* (Vol. 1, pp. 459–464). IEEE.
[14] Zhu, F., Zhu, C., Fang, Z., Lu, W., & Pan, J. (2025). Using constrained K-means clustering for soil texture mapping with limited soil samples. *Agronomy, 15*(5), 1220.
[15] Bougiouklis, J.-N., Barouchas, P. E., Petropoulos, P., Tsesmelis, D. E., & Moustakas, N. (2025). Precision soil sampling strategy for the delineation of management zones in olive cultivation using unsupervised machine learning methods. *Scientific Reports, 15*(1), 8253.
[16] Gulhane, V. A., Rode, S. V., & Pande, C. B. (2023). Correlation analysis of soil nutrients and prediction model through ISO cluster unsupervised classification with multispectral data. *Multimedia Tools and Applications, 82*(2), 2165–2184.
[17] Zhao, W., Unagaev, A., & Efremova, N. (2025). Data-driven soil organic carbon sampling: Integrating spectral clustering with conditioned Latin hypercube optimization. *arXiv*, arXiv:2506.10419.
[18] Novielli, P., Magarelli, M., Romano, D., de Trizio, L., Di Bitonto, P., Monaco, A., Amoroso, N., et al. (2024). Climate change and soil health: Explainable artificial intelligence reveals microbiome response to warming. *Machine Learning and Knowledge Extraction, 6*(3), 1564–1578.
[19] Swain, K. P., Nayak, S. R., Ravi, V., Mishra, S., Alahmadi, T. J., Singh, P., & Diwakar, M. (2024). Empowering crop selection with ensemble learning and K-means clustering: A modern agricultural perspective. *The Open Agriculture Journal, 18*(1).
[20] Muhammed, D., Ahvar, E., Ahvar, S., & Trocan, M. (2023). Performance evaluation of machine learning algorithms for a cluster-based crop recommendation system. In *2023 17th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)* (pp. 441–445). IEEE.
[21] Wang, H., Zhao, Y., Li, S., Liu, Z., & Zhang, X. (2025). DeepCropClustering: A deep unsupervised clustering approach by adopting nearest and farthest neighbors for crop mapping. *ISPRS Journal of Photogrammetry and Remote Sensing, 224*, 187–201.
[22] Kulatunga, C., Dhelim, S., & Kechadi, T. (2024). Machine learning for dynamic management zone in smart

farming. *arXiv*, arXiv:2408.00789.
[23] Chaali, N., Ramírez-Gómez, C. M., Jaramillo-Barrios, C. I., Garré, S., Barrero, O., Ouazaa, S., & Calderon Carvajal, J. E. (2024). Enhancing irrigation management: Unsupervised machine learning coupled with geophysical and multispectral data for informed decision-making in rice production. *Smart Agricultural Technology, 9*, 100635.
[24] Prity, F. S., Hasan, M. M., Saif, S. H., Hossain, M. M., Bhuiyan, S. H., Islam, M. A., & Lavlu, M. T. H. (2024). Enhancing agricultural productivity: A machine learning approach to crop recommendations. *Human-Centric Intelligent Systems*, 1–14.
[25] Zhao, K., Wu, S., Liu, C., Wu, Y., & Efremova, N. (2023). Precision agriculture: Crop mapping using machine learning and Sentinel-2 satellite imagery. *arXiv*, arXiv:2403.09651.
[26] Nagy, J., Zalai, M., Illés, Á., & Monoki, S. (2024). The impact of crop year and crop density on the production of sunflower in site-specific precision farming in Hungary. *Agriculture, 14*(9), 1515.