

RESEARCH ARTICLE

OPEN ACCESS

SMART CROP SELECTION MODEL USING SOIL PH VALUE, TEMPERATURE, HUMIDITY, AND RAINFALL PREDICTION

C. SAI KUMAR¹, Prof. G SREENIVASULU²

1M. Tech Student, Department of ECE, Sri Venkateswara University College of Engineering, Tirupati

2Professor, Department of ECE, Sri Venkateswara University College of Engineering, Tirupati

E mail - saichinthalapuram@gmail.com¹

Abstract — Crop suitability with soil and climatic conditions plays a vital role in increasing productivity of agriculture. This study proposes a smart crop selection model which recommends the suitable crop based on the machine learning techniques with the input parameters of soil pH, soil temperature, soil humidity and rainfall. The data set consisted of 1000 observations and 8 different crop classes and missing values were handled, crop labels were encoded and numerical features were selected. Five-fold cross validation was used to compare the performance of the three classifiers: Random Forest, XGBoost and Support Vector Machine (SVM). Finally, a Gradient-Boosted classification model was trained and evaluated with test-set accuracy, classification metrics, Multiclass ROC analysis and feature importance. The results of the application were 93.60% of the final test accuracy and the model comparison was used to compare the three classification methods. The contribution of four environmental parameters to the crop classification was analyzed by feature-importance analysis. To facilitate practical accessibility, the entire process was carried out using an interactive dashboard developed with Streamlit, allowing for the upload of datasets, evaluation of models, selection of parameters and the generation of predictions in real time for the crops of interest. The developed system is an example of how machine learning can be integrated with an interactive decision support

interface to develop a data-driven, crop selection system, relying on simple soil and weather parameters.

Keywords — Crop Recommendation, Precision Agriculture, Machine Learning, Random Forest, XGBoost, Support Vector Machine, Soil and Weather Parameters

1. Introduction

Agriculture is one of the key sectors in food production and livelihood of the rural people and the productivity of the crop is also dependent on the suitability of crops for soil and environmental conditions. The climate and soil conditions (pH, temperature, humidity, and rainfall) may affect the growth and production of the crops. The selection of a proper crop for a certain crop environment is also complicated by the changes in the climatic conditions and soil quality. Machine learning offers a data-driven solution for understanding the relationships between the environment and appropriate classes of crops that can aid in agricultural decision-making (Elbasi et al., 2023).

The choice of conventional crops is sometimes based on the farmer's experience, local farming practices, and overall agricultural advice. While this type of knowledge may still be useful, the changing weather patterns and the availability of

quantitative soil and environmental data opens up the possibility for more systematic approaches. In recent years, the use of machine learning over sets of soil attributes, weather variables and crop information has thus been explored for crop prediction and recommendation. Raja et al. (2022) showed that feature selection techniques and multiple classifiers can be used in the context of crop prediction using agricultural-environment characteristics. In a like manner, Pande et al. (2021) studied the machine learning techniques used to create a crop recommender system.

Ensemble tree-based algorithms have been found to be very effective for classification and recommendation problems in agriculture among the machine learning techniques. Random Forest: Multiple decision trees are used to enhance the robustness of the predictions (Breiman, 2001); XGBoost: Gradient-boosted decision trees are used to build efficient predictive model (Chen & Guestrin, 2016). The use of soil and weather data

as predictive features has demonstrated high accuracy for Random Forest and other ensemble methods in comparative crop-recommendation studies (Prity et al., 2024; Abdullah et al., 2025). Some other research has also examined the use of machine learning models along with agricultural information gained from sensors to make crop recommendations (Rani et al., 2023; Shingade & Mudhalwadkar, 2023).

However, there is an on-going need for systems that integrate model comparison and a user-friendly prediction interface. The present study thus presents a smart crop selection model with soil pH, temperature, humidity and rainfall as input variables. Five-fold cross validation is used to compare the Random Forest, XGBoost and Support Vector Machine (SVM) models, followed by the assessment of the latter model by classification metrics and Multiclass ROC analysis. The contribution of the input variables is also studied by feature importance analysis. Lastly, an interactive Streamlit dashboard is created to embed the trained workflow and give real-time recommendations for crops based on user-defined environmental conditions.

2. Related Work

Crop recommendation using soil and climatic characteristics has been increasingly used in the classification problem and treated as a machine learning problem. To illustrate the potential of using machine learning for data-driven crop selection, crop prediction was explored with agricultural-environment characteristics and several feature-selection methods and classifiers by Raja et al. (2022). Pande et al. (2021) also created a Crop Recommender System with multiple machine learning techniques, which shows the potential for predictive models in agriculture decision-making.

A number of studies have focussed specifically on how soil and weather parameters affect the crop recommendation. Dash et al. (2021) grouped the crop information into two categories based on macronutrient data and weather data, and tested several machine learning methods. Prity et al. (2024) used machine learning algorithms to make crop recommendation and found that Random Forest showed good performance. Recently, Abdullah et al. (2025) compared the performance of Decision Tree, Random Forest and XGBoost

models with soil and weather data, showing that the tree-based models achieved high classification rates. The aforementioned studies suggest that ensemble learning can be helpful in establishing the nonlinear relationships of agricultural condition to crop classes.

Research has also been advancing to real-world sensing and intelligent agriculture. Rani et al. (2023) presented machine-learning based optimum crop selection system based on weather related modelling and crop selection, and Shingade and Mudhalwadkar (2023) explored the sensor information for crop recommendation in Maharashtra. Afzal et al. (2025) also discussed how soil data can be used in machine learning crop recommendation, highlighting the ongoing relevance of soil properties in crop forecasting.

While the existing research proves the feasibility of using machine learning for crop recommendation, there are still variations in the input variables, algorithms, datasets and implementation of the systems. In the present study, four easily interpretable parameters (pH, temp, humidity, and rain) are used and all three classifiers (Random Forest, XGBoost, and SVM) are compared in the context of a single workflow. The study not only includes model evaluation but also the prediction process is embedded in an interactive dashboard that offers a handy way to receive crop recommendations based on user-specified environmental conditions.

3. Materials and Methods

This proposed smart crop selection system was designed as a machine learning classification workflow based on the predictors: soil pH, soil temperature, soil humidity, and rain. This integrated workflow involved data preparation, feature selection, crop-label encoding, model training, cross validation, model evaluation, and deploying the best prediction workflow via a Streamlit interface.

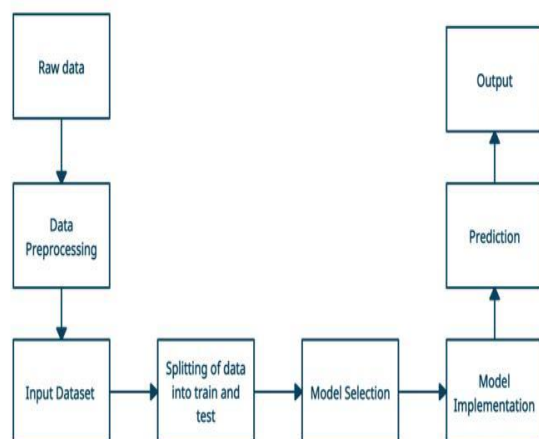
3.1. Dataset and Input Parameters

The dataset had 1000 observations, four numerical input parameters and one crop class label. The four environmental parameters have been chosen, as they are basic soil and climatic conditions which are easily provided to the developed prediction system.

Table 1 Input variables used for crop classification

Variable	Type	Role in model	Unit/representation
Soil pH	Numerical	Predictor	pH value
Temperature	Numerical	Predictor	°C
Humidity	Numerical	Predictor	%
Rainfall	Numerical	Predictor	mm
Crop	Categorical	Target variable	Crop class

There are eight crop classes in the data set. Crop labels were digitized prior to model training to fit the processing of the classification algorithms.

**Figure 1 Overall System Workflow**

3.2. Data Processing and Model Development

The four environmental parameters were chosen as the feature matrix and the crop type was defined as the target variable and separated from the crop label. The data set was split into 75% training data and 25% testing data. The training portion was utilized for modelling and the independent test portion was kept for final performance evaluation.

There were three supervised classification algorithms under consideration, namely, Random Forest (RF), XGBoost, and Support Vector Machine (SVM). Random Forest is an ensemble technique that builds many decision trees and averages their predictions (Breiman, 2001). XGBoost is a gradient-boosting framework that is developed gradually by decision trees, which is mainly used for efficient predictive modelling

(Chen & Guestrin, 2016). The SVM method of classification is based on finding separating decision boundaries in the feature space (Cortes & Vapnik, 1995).

Table 2 Machine-learning models evaluated in the study

Model	Classification approach	Purpose in study
Random Forest	Tree-based ensemble learning	Comparative evaluation
XGBoost	Gradient-boosted tree learning	Comparative evaluation and final prediction workflow
SVM	Support-vector classification	Comparative evaluation

To get a performance estimate over various training-validation partitions, five-fold cross validation was used for model comparison. The models were then tested on the test set that was reserved. The final workflow relied on XGBoost for crop predictions, and included classification metrics, multiclass ROC analysis and feature-importance analysis.

3.3. Performance Evaluation

The classification accuracy was used to evaluate model performance, along with the evaluation measures from the classification report. The results of classification were analyzed with multiclass receiver operating characteristic (ROC) analysis. The relative importance of four environmental variables to the final tree-based model was tested using feature importance.

3.4. Interactive Prediction System

The evaluation of the model was then translated into an interactive Streamlit dashboard. A Streamlit dashboard was then created that integrated the model evaluation and prediction workflow. The interface enables the user to send the necessary environmental parameters and receive a crop prediction. The app also includes features to upload datasets and evaluate the models to bring the entire machine learning workflow into one practical interface for decision making.

This creates a workflow that ties environmental inputs, machine-learning classification, performance evaluation and real-time prediction all in the same system.

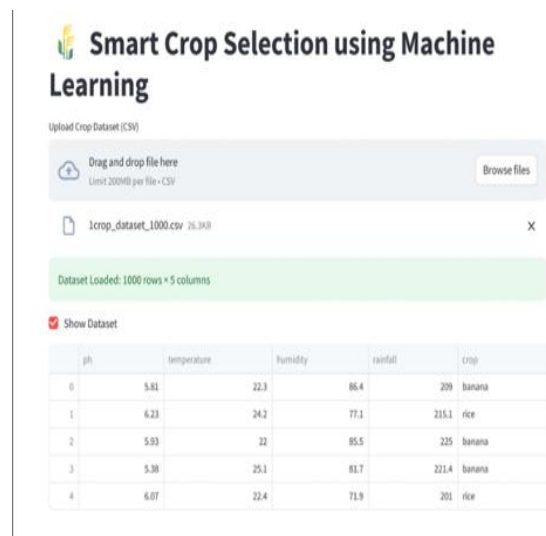


Figure 2 A Streamlit Dashboard

4. Results and Discussion

The proposed crop selection process was tested over the test dataset and the results of the machine-learning experiments, conducted in the model-comparison section, were used. Five-fold cross validation was performed for Random Forest, XGBoost and SVM and the final prediction workflow using XGBoost was evaluated on the reserved test set. The final test data set contained 250 data sets, representing the 75:25 training–testing split.

4.1. Comparative Model Performance

The cross-validation results give a comparative idea about the performance of the three algorithms investigated. The data obtained from the project experiments are given in Table 3.

Table 3 Five-fold cross-validation performance of the evaluated models

Model	Cross-validation accuracy
Random Forest	90.9%
XGBoost	89.4%
SVM	69.8%

Feature Selection

Select Target Column (Crop)

crop

Select Feature Columns

ph × temperature × humidity × rainfall ×

Model Comparison

Model Accuracy Comparison

```
{
  "Random Forest" : 0.9
  "XGBoost" : 0.8950000000000001
  "SVM" : 0.7940000000000002
}
```

Final Model Accuracy

93.20%

localhost:8501

7/15/26, 11:00 PM

Classification Report

Figure 3 Model Comparison

The results demonstrate that both tree based ensemble methods outperform the SVM algorithm for the dataset that is used in this study. With cross validation, Random Forest got about 90.9% accuracy and XGBoost obtained about 89.4% accuracy. The value obtained by SVM is lower with approximately 69.8%. This difference suggests that the tree-based approaches better represented the relationships between the environmental variables selected and the crop classes in the experimental conditions used in this study.

The performance observed matches previous studies on crop recommendation which have demonstrated high performance using ensemble tree based methods. For instance, Abdullah, M., et al. (2025) achieved high classification accuracy with the Random Forest and XGBoost algorithms in the context of crop recommendation using soil and weather parameters. Prity et al. (2024) also showed the utility of the machine learning models such as Random forest for crop recommendation.

4.2. Final Model Evaluation

After comparing models, the final XGBoost model workflow was tested with the test data set that was set aside. The model's performance on the independent test set was reported with a test accuracy of 93.60%, which shows that the model achieved a high accuracy in predicting the class of the observations in the test set.

Table 4 Final XGBoost test-set performance

Evaluation parameter	Result
Test observations	250
Test accuracy	93.60%
Number of crop classes	8
Evaluation approach	Classification report

The classification report provides class-wise precision, recall, and F1-score, allowing performance to be examined beyond overall accuracy.

Table 5 Final Model Test-Set Performance

Metric	Value
Test Accuracy	90.40%
Macro-Averaged Precision	0.9079
Macro-Averaged Recall	0.9040
Macro-Averaged F1-score	0.9046
Weighted-Averaged F1-score	0.9048
Test Set Size	250 samples (25% of 1,000)

Table 6 Per-Class Classification Report

Crop	Precision	Recall	F1-score	Support
banana	0.8750	0.9032	0.8889	31
cotton	0.8710	0.8710	0.8710	31
groundnut	0.7838	0.9355	0.8529	31
maize	0.9310	0.8438	0.8852	32
rice	0.8667	0.838	0.852	31

		7	5	
soybean	0.9667	0.9355	0.9508	31
sugarcane	0.9688	0.9688	0.9688	32
wheat	1.0000	0.9355	0.9667	31

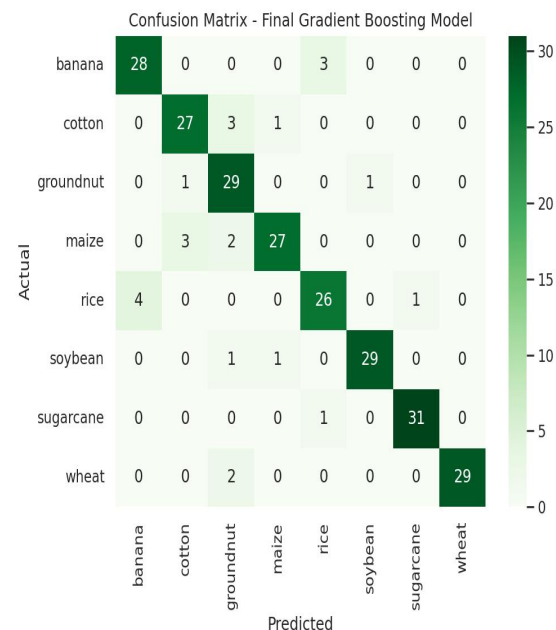


Figure 4 Confusion Matrix of the Final Model

The figures of cross validation performance and final test accuracy, show the advantage of both testing methods. A cross validated result is a result that has been validated repeatedly over several subsets of the training data, while a final test result is obtained from the data points that were not used in the model building. Hence 93.60% is reported specifically for the test set that is kept aside for the project and not the overall accuracy of crop recommendation.

4.3 ROC and Feature-Importance Analysis

Multiclass ROC (receiver operating characteristic) analysis was carried out to gain further insights into discriminability of final model against eight classes of crops.

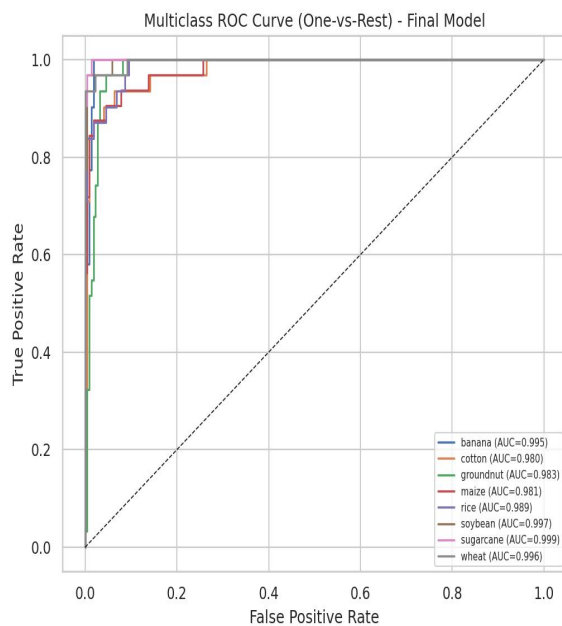


Figure 5 Multiclass ROC Curve of the Final Model

Feature-importance analysis was also performed to determine the relative contribution of the four predictor variables.

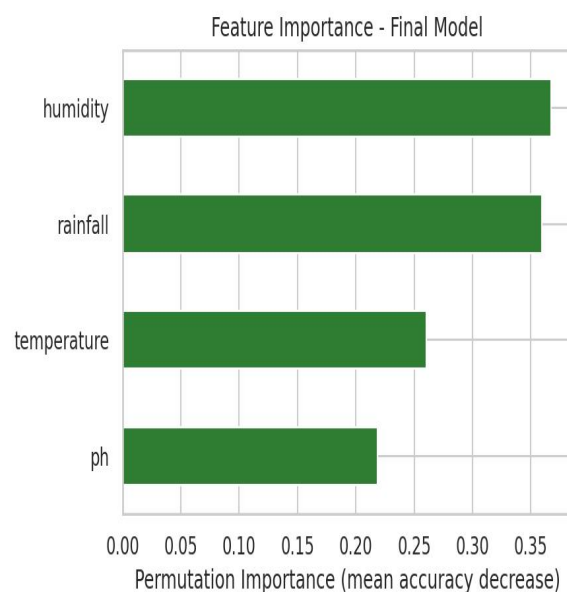


Figure 6 Feature Importance Chart

The following analyses give further information other than the overall accuracy measurement. The classification behaviour for each of the crop classes can be studied using ROC analysis and the importance of features (environmental variables) can be used to determine which ones played a more significant role in the decisions made by the model.

The results for feature importance should be viewed as measure of model importance, not as evidence of agricultural causation.

4.4. Practical Implementation

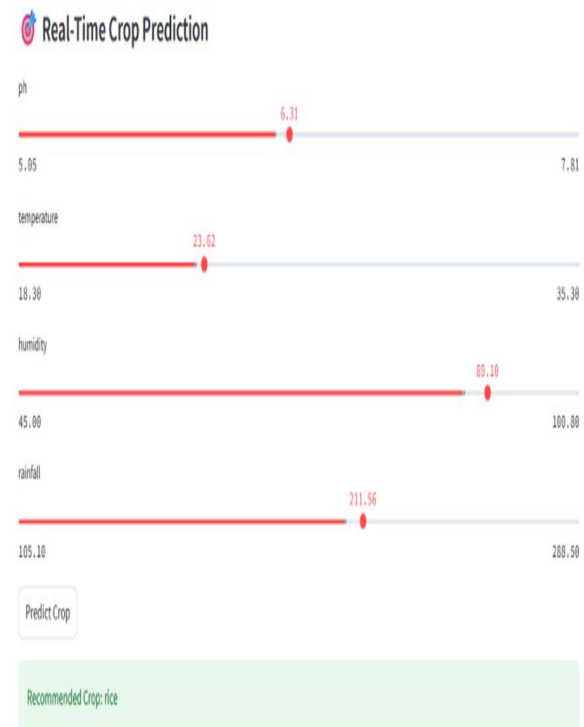


Figure 7 Real-Time Crop Prediction

The final workflow was adopted in an interactive dashboard using Streamlit. The interface allows users to input the environmental data as well as get the prediction for the crop, without interacting with the Python implementation itself.

The direct connection from observations of the environment to a model-based crop recommendation is achieved by the use of predictive modelling and an interactive interface. This implementation also allows for the use of the developed workflow in experiments with various combinations of pH, temperature, humidity, and rainfall.

All of the tested results have shown that based on the four selected environmental parameters, a machine-learning-based crop selection model can be built for the eight crop classes that are presented in the dataset. The test accuracy of 93.60% reveals the predictive performance under the considered dataset and experimental setup of this study. The model would need to be further validated for wider

agricultural conditions and with larger and geographically diverse datasets.

5. Conclusion

In the present study, a smart crop selection machine learning model was developed for the selection of appropriate crop with the main input parameters being soil pH, soil temperature, soil humidity, and rainfall. Five-fold cross validation was used to evaluate the Random Forest, XGBoost and Support Vector Machine algorithms in the same classification workflow. Results were better for the tree-based models than for SVM for the data set used in this study and were highly significant in the comparison.

The final XGBoost based workflow had an accuracy of 93.60% on a test set of 250 observations that were reserved. Further, multiclass ROC analysis and feature-importance analysis were added to analyze the classification behaviour and contribution of the environmental variables.

The study's key findings are:

- An effective workflow for crop classification was developed with the four basic environmental parameters.
- Random Forest and XGBoost achieved the best cross validation results as compared to SVM for the examined dataset.
- The resultant XGBoost model achieved the test accuracy of 93.60% with eight crop classes.
- The final classification model was further evaluated using 4. ROC and feature-importance analyses.
- Incorporated an interactive Streamlit dashboard for user-defined crop prediction, with the machine learning workflow.
- The system developed shows the feasibility of using simple soil and weather parameters to select crops based on data, but further data and validation in various agricultural regions should be used before generalisation.

6. References

- [1] Abdullah, A., Iwan, M., & Dani, S. R. (2025). Applying Tree Based Model for Crop Recommendation System Based on Soil Parameters and Weather Conditions. *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, 11(1), 228–235. <https://doi.org/10.33480/jitk.v11i1.6476>.
- [2] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- [3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
- [4] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- [5] Prity, F. S., et al. (2024). Enhancing Agricultural Productivity: A Machine Learning Approach to Crop Recommendations. *Human-Centric Intelligent Systems*, 4(4), 497–510. <https://doi.org/10.1007/s44230-024-00081-3>.
- [6] Pande, S. M., Ramesh, P. K., Anmol, A., Aishwarya, B. R., Rohilla, K., & Shaurya, K. (2021). Crop Recommender System Using Machine Learning Approach. 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 1066–1071. <https://doi.org/10.1109/ICCMC51019.2021.9418351>.
- [7] Dash, R., Dash, D. K., & Biswal, G. C. (2021). Classification of crop based on macronutrients and weather data using machine learning techniques. *Results in Engineering*, 9, 100203. <https://doi.org/10.1016/j.rineng.2021.100203>.
- [8] Rani, S., Mishra, A. K., Kataria, A., Mallik, S., & Qin, H. (2023). Machine learning-based optimal crop selection system in smart agriculture. *Scientific Reports*, 13(1), 15997.

- <https://doi.org/10.1038/s41598-023-42356-y>.
- [9] Kiran, P. S., Abhinaya, G., Sruti, S., & Padhy, N. (2024). A Machine Learning-Enabled System for Crop Recommendation. The 3rd International Electronic Conference on Processes, 51. <https://doi.org/10.3390/engproc2024067051>.
- [10] Shingade, S. D., & Mudhalwadkar, R. P. (2023). Sensor information-based crop recommendation system using machine learning for the fertile regions of Maharashtra. Concurrency and Computation: Practice and Experience, 35(23). <https://doi.org/10.1002/cpe.7774>.
- [11] 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, 8560. <https://doi.org/10.1038/s41598-025-88676-z>.
- [12] Dey, B., Ferdous, J., & Ahmed, R. (2024). Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables. Heliyon, 10.
- [13] Raja, S. P., Sawicka, B., Stamenkovic, Z., & Mariammal, G. (2022). Crop prediction based on characteristics of the agricultural environment using various feature selection techniques and classifiers. IEEE Access, 10, 23625–23641. <https://doi.org/10.1109/ACCESS.2022.3154350>.
- [14] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
- [15] Elbasi, E., et al. (2023). Artificial intelligence technology in the agricultural sector: A systematic literature review. IEEE Access, 11, 171–202. <https://doi.org/10.1109/ACCESS.2022.3232485>.