Inter Crop Recommendation using Machine Learning and Rule based system

CS4099 Project Final Report

Submitted by

Pitta Partha Saradhi Reddy	**B210582CS**
Bypuri Sai Narayana Koushik	**B210567CS**
Pagolu Rakshith Pradeep	**B210576CS**

Under the Guidance of
Dr. Anil Pinapati

Department of Computer Science and Engineering
National Institute of Technology Calicut
Calicut, Kerala, India - 673 601

April 15, 2025

NATIONAL INSTITUTE OF TECHNOLOGY
CALICUT, KERALA, INDIA - 673 601

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

तमसो मा ज्योतिर्गमय

2025

CERTIFICATE

Certified that this is a bonafide record of the project work titled

PROJECT TITLE

done by

Pitta Partha Saradhi Reddy
Bypuri Sai Narayana Koushik
Pagolu Rakshith Pradeep

of eighth semester B. Tech in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering of the National Institute of Technology Calicut

Project Guide	**Head of Department**
Dr. Anil Pinapati	Dr. Subashini R
Assistant Professor	Associate Professor

DECLARATION

We hereby declare that the project titled, **Inter Crop Recommendation using Machine Learning and Rule based system**, is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or any other institute of higher learning, except where due acknowledgement and reference has been made in the text.

Place : NIT Calicut
Date : 14-04-2025

Signature :
Name : PITTA PARTHA SARADHI REDDY
Reg. No. : B210582CS

Abstract

This project aims to help farmers choose better crop combinations using a two-step system. First, we use machine learning to recommend the best crop (Crop-1) to grow based on soil and weather conditions like temperature, humidity, pH, rainfall, and nutrient levels (N, P, K). Once the main crop is selected, we use a rule-based system to suggest other crop (Crop-2) that can be grown alongside it. These recommendations are based on how well the crops match in terms of water and nutrient needs, root depth, pest and disease resistance, maturity time, shade tolerance, and season. By combining data-driven predictions with expert rules, our system helps support smarter and more sustainable inter-cropping decisions.

ACKNOWLEDGEMENT

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Choosing the right crops to grow is an important decision for farmers, especially when trying to use their land more efficiently. Inter cropping—growing two or more crops together—can increase productivity, improve soil quality, and reduce pests. But picking the right crops to grow together isn't easy, as it depends on many factors like the season, soil conditions, and how different crops interact with each other.

In this project, we've built a two-part system to make this easier. First, we use machine learning to look at things like soil nutrients, temperature, humidity, and rainfall to recommend the best main crop (Crop-1) for the current conditions. Then, using that crop and the season as input, we apply a rule-based system to recommend two compatible inter crops (Crop-2). These suggestions are based on agricultural knowledge, taking into account water and nutrient needs, root depth, pests and diseases, shade tolerance, and the time each crop takes to grow. This approach helps farmers make better choices and supports sustainable farming..

Chapter 2

Literature Survey

Agriculture is a vital sector that significantly impacts the economy and food security worldwide. With the increasing global population and unpredictable climatic conditions, ac- curately predicting crop yields has become more challenging yet essential for optimizing agricultural practices and ensuring food supply. Traditional methods for crop yield prediction often rely on statistical models, which may not effectively handle the complexity of factors such as weather patterns, soil conditions, and crop characteristics. In recent years, ma- chine learning (ML) and deep learning (DL) techniques have emerged as powerful tools for improving prediction accuracy by capturing non-linear relationships and time-dependent pat- terns in agricultural data.

One study presented a methodology for crop yield pre- diction using modular artificial neural networks (MANNs) and support vector regression (SVR) outperforms traditional methods [4]. The study focused on estimating crop yields for the Kharif season in the Vishakhapatnam district of Andhra Pradesh. Rainfall significantly influences Kharif crops pro- duction; hence, they used modular artificial neural networks to forecast rainfall before using support vector regression to estimate crop yield using rainfall and area data. Machine learning (ML) is transforming agriculture by mak- ing it more efficient and productive.This study explain how ML is used in areas like monitoring crop health, predicting yields, managing soil and water, and detecting pests and weeds. Technologies such as IoT sensors, drones, and satellites collect data that ML models process to help farmers make better decisions. However, there are still challenges like high costs, data compatibility, and security concerns, making it important to develop more affordable and practical solutions for farmers worldwide [1].

The best model for crop yield prediction was identified by a comparison of algorithms such as K-Nearest Neighbor, Decision Tree, and Random Forest based on soil nutrients and climatic factors using a dataset of 22 crops [3]. The study compared the models under two criteria — Entropy and Gini Index and found that Random Forest achieved the highest accuracy of 99.32 %, outperforming KNN and Decision Tree. This research emphasizes the effectiveness of Random Forest in crop prediction and suggests its potential to help farmers make informed decisions about crop selection. In a study machine learning techniques were used to predict crop yield based on features such as temperature, rainfall, humidity, pH levels, and crop types.

The Random Forest al gorithm was applied and compared with models like Decision Trees and Support Vector Regression (SVR). Random Forest provided the highest accuracy, demonstrating its effectiveness in predicting crop yields. This approach highlights the im- portance of using environmental factors to help farmers make informed decisions about crop selection [2]. Intercropping has gained widespread recognition as a sus- tainable alternative to monocropping due to its many benefits for agricultural productivity, soil health, and resource effi- ciency. According to Gebru (2015), intercropping stabilizes crop yields and maximizes land use efficiency by leveraging the complementary growth characteristics of different crops. Research indicates that intercropping systems can utilize sun- light more effectively, thanks to the varying heights and root structures of the crops involved. In addition, inter-cropping aids in pest control, reduces soil degradation, and offers economic advantages through the diversification of farmers' income. For developing countries, inter-cropping serves as an effective strategy to improve food security while minimizing dependence on chemical inputs. These insights provide a solid foundation for incorporating intercropping into modern machine learning-based yield prediction models, as explored in this study, to improve decision making and promote agronomic sustainability. [5]

Chapter 3

Problem Definition

Farmers often struggle to choose the right crops for inter-cropping, as it requires careful planning to ensure compatibility between crops. Most existing systems only recommend a single crop based on soil and weather conditions, but they do not help in selecting suitable inter crops.

Choosing incompatible crops can lead to poor yield, competition for resources, and pest issues. To solve this, there is a need for a system that can first recommend the best main crop using machine learning, and then suggest two compatible inter crops based on agricultural rules and seasonal factors.

This project aims to develop a two-stage system that helps farmers make better decisions for inter cropping by combining data-driven recommendations with domain knowledge.

Chapter 4

Methodology

This project aims to recommend optimal inter cropping systems using a two-stage approach:

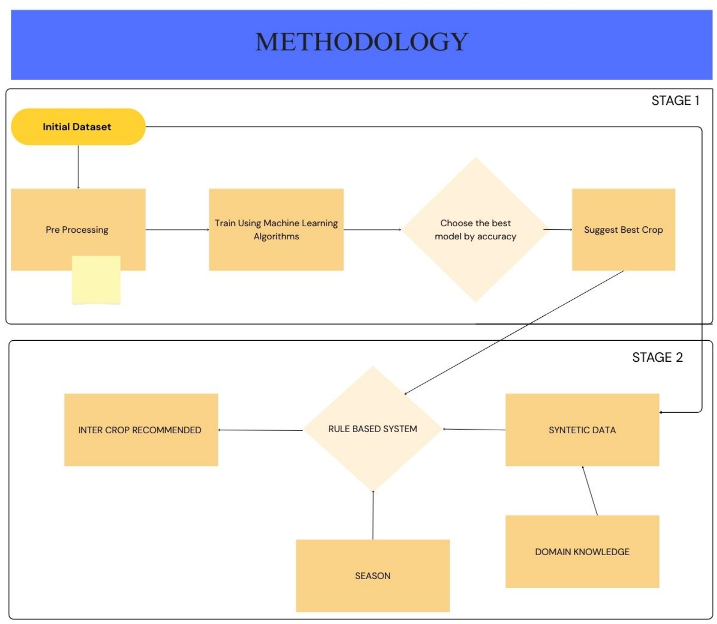

Figure 4.1: Proposed Methodology

1. Primary crop prediction using machine learning algorithms based on environmental and soil parameters.

2. Rule-based intercrop recommendation system that selects suitable companion crops based on compatibility criteria.

The methodology proposed for our project can be seen in Figure 4.1

Stage 1: Primary Crop Prediction using Machine Learning

A.Dataset Description:
We used the Crop Recommendation Dataset[6] which includes soil and environmental parameters as input characteristics and a target label of the crop name. The Features are shown in Table 4.1

Feature	**Description**
`N`, `P`, `K`	Nitrogen, Phosphorus, and Potassium content in the soil
`temperature`	Average temperature in degrees Celsius
`humidity`	Relative humidity in percentage (%)
`ph`	Soil pH level
`rainfall`	Annual rainfall
`label`	Target crop name

Table 4.1: Input features in the crop recommendation dataset

We have trained this dataset using multiple Machine Learning algorithms:

B. Machine Learning Methods:

1. **Linear Regression:**
   Linear Regression is a foundational supervised learning algorithm used to model the relationship between a dependent variable and one or

more independent variables. It assumes a linear relationship between input features and the output.

- Justification: In our project, Linear Regression can be used for tasks like predicting expected crop yield based on soil nutrients and environmental conditions. It fits a linear model to the training data to estimate continuous output values.
- Benefits: Easy to interpret, fast to train, and useful as a baseline regression model. It helps in understanding the impact of each feature on the output variable.
- Use Case in Our Project: Ideal when estimating continuous values like expected crop yield, nutrient requirement, or growth time. It performs well when the relationship between features and output is approximately linear.

2. **Logistic Regression:**
Logistic Regression is a statistical method used for classification problems, particularly effective for binary and multi-class scenarios.

- Justification: In our case, LR is applied to predict the most suitable crop (multi-class classification) based on continuous environmental and soil features. Logistic Regression creates a linear decision boundary using a logistic (sigmoid) function that maps input features to output class probabilities.
- Benefits: It is simple to implement, computationally efficient, and interpretable. LR performs well when there is a near-linear relationship between the independent features (e.g., nitrogen, pH, humidity) and the log-odds of the crop classes.
- Use Case in Our Project: Logistic Regression was used as a baseline model for comparison against more complex classifiers like Random Forest and SVM. It is particularly useful for small to

medium-sized datasets where the decision surface is expected to be linear.

3. **Support Vector Machine**
   Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification. It aims to find the optimal hyperplane that maximally separates data points of different classes.

   - Justification: In our project, SVM is used to classify the best crop based on input features (N, P, K, temperature, humidity, pH, rainfall). The algorithm attempts to create the largest possible margin between crop classes for improved generalization.
   - Benefits: SVM is highly effective in high-dimensional spaces and can be extended with kernel functions to handle non-linear classification. It is also resistant to overfitting, especially in cases where a clear margin exists between crop classes.
   - Use Case in Our Project: Suitable for structured datasets with clear crop-class boundaries. With kernel tricks like RBF or polynomial, it performs well even when the data isn't linearly separable.
     The decision boundary is defined by a hyperplane:

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{4.1}$$

     Where:

     – w is the weight vector,
     – x is the input feature vector,
     – b is the bias.

The margin is maximized by:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{subject to } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad \forall i \tag{4.2}$$

4. **K-Nearest Neighbors**

   The K-Nearest Neighbors (KNN) algorithm is a non-parametric, instance-based learning method used for classification. It assigns a class to a new input based on the majority class of its k nearest neighbors in the feature space.

   - Justification: In our project, KNN is used to classify the most suitable crop by comparing the given environmental and soil conditions with similar instances in the dataset. It relies on feature similarity to make predictions without any prior training.
   - Benefits: KNN is easy to implement, intuitive, and effective for small datasets. It does not involve a training phase, making it computationally efficient for applications with moderate-sized datasets. It also adapts well to non-linear decision boundaries.
   - Use Case in Our Project: KNN is used to find the crop most frequently grown under similar environmental conditions (based on N, P, K, temperature, humidity, pH, and rainfall). It acts as a strong baseline and performs well when the dataset is clean and moderately sized.

5. **Decision Tree**

   The Decision Tree Classifier is a non-parametric supervised learning method used for classification tasks. It works by learning decision rules inferred from the data features.The Decision Tree model (see Figure 4.2) uses a flowchart-like structure to make decisions based on feature values, leading to a classification at the leaf nodes.

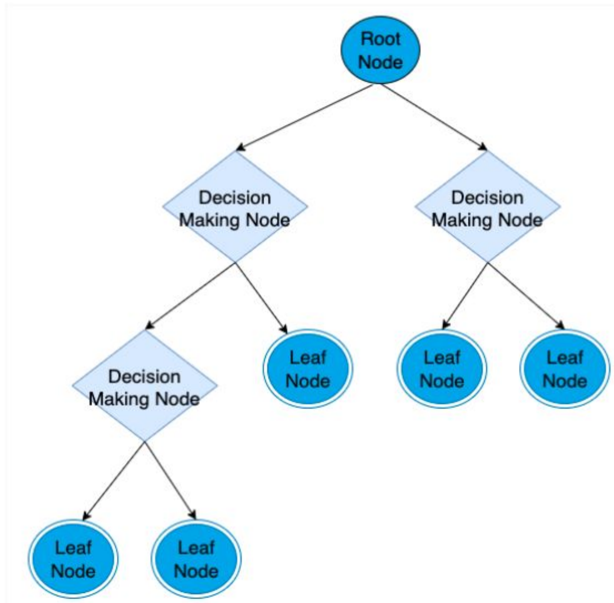

Figure 4.2: Architecture of a Decision Tree

- Justification: In our project, a Decision Tree is used to predict the most suitable crop by recursively splitting the dataset based on feature thresholds (e.g., N, P, K, temperature, etc.). The tree structure simulates a sequence of decisions leading to the final crop label.
- Benefits: It is easy to interpret, does not require feature scaling or normalization, and naturally handles both numerical and categorical data. The model highlights important features based on their use in decision splits.
- Use Case in Our Project: Decision Trees are especially useful for identifying key soil or environmental factors influencing crop selection. Their transparent nature helps in understanding the decision-making process and improving trust in model recommendations.

6. **Random Forest**

Random Forest is an ensemble learning algorithm widely used for classification and regression tasks. It builds multiple decision trees on random subsets of the training data and aggregates their outputs to improve overall accuracy and robustness.The Random Forest model (Figure 4.3) combines multiple such trees and predicts the class based on a majority vote, thus improving accuracy and reducing overfitting.

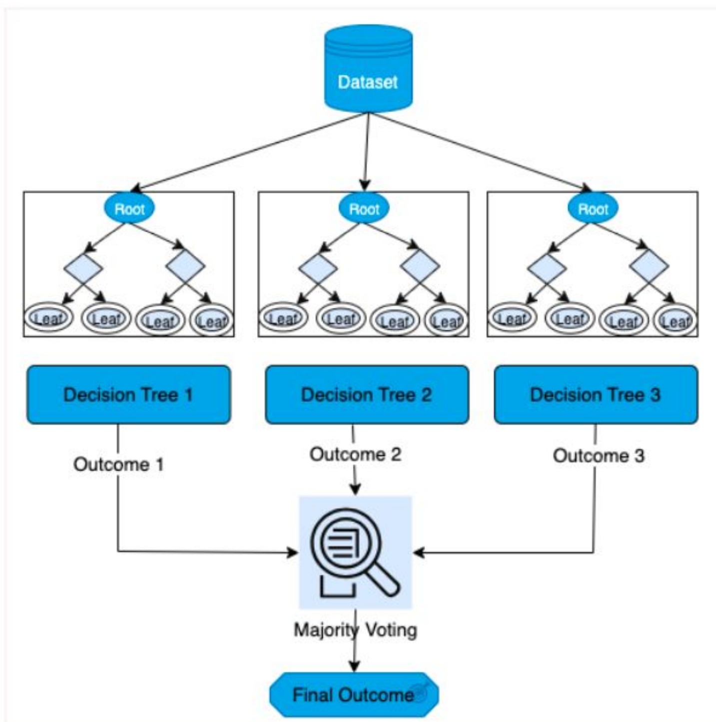

Figure 4.3: Random Forest model using multiple decision trees

- Justification: In our project, Random Forest is used to predict the most appropriate crop based on input features such as soil nutrients, temperature, pH, and rainfall. By combining the outputs of several decision trees, it reduces overfitting and provides more stable and accurate predictions compared to a single tree.

- Benefits: Random Forest handles both numerical and categorical data effectively, is resistant to overfitting, and can work well with datasets that include noise or missing values. Additionally, it provides insights into feature importance, helping identify which environmental factors most strongly influence crop suitability.
- Use Case in Our Project: Random Forest serves as one of the most effective classifiers for crop recommendation, especially when the dataset has complex interactions and patterns. It has been found to outperform simpler models in terms of accuracy and generalization.

C. Choosing the Best Accuracy Model:

In this stage, multiple machine learning models were developed to identify the most suitable algorithm for crop prediction based on input features. The data set used includes essential agricultural parameters such as Nitrogen (N), Phosphorus (P), Potassium (K), Temperature, Humidity, pH level, and rainfall. The target variable represents the crop to be cultivated under given conditions.The models described in the above section are considered. Each model was trained using the same preprocessed dataset. Preprocessing steps included handling missing values, scaling the features if needed, and splitting the data into training and testing sets. Hyperparameter tuning and model fitting were performed individually for each algorithm.

After training, these models were prepared for evaluation to determine which algorithm aligns best with the classification task. The model selected from this process was then used in the first stage of the crop recommendation system and the primary crop is resulted which will then be used later in stage 2.

Stage 2 – Intercrop Recommendation Using Rule-Based System

D.Synthetic Data Generation:

To build an effective rule-based system for intercrop recommendation, a dataset containing various compatibility features between potential crop pairs was needed. Since such detailed data is not readily available in standard datasets, synthetic data generation was performed. This process involved combining known data, agronomic principles, and expert knowledge to create a realistic and usable dataset. The features of the Synthetic data are shown in Table 4.2

- Water and Nutrient Requirements for both crops: These values for the primary crop (crop-1) and intercrop(crop-2) are derived from the original crop dataset used in Stage 1. For intercrops (crop-2), that are not in dataset 1 the values are inferred or assigned based on domain expertise and known agronomic behavior
- Root Depth Levels: Indicates whether a crop has shallow, medium, or deep roots. Used to avoid underground resource competition.
- Pest and Disease Overlap (%): Represents the percentage of pests or diseases shared between the two crops. Lower values are preferred.
- Days-to-Maturity Difference: Helps ensure the crops have compatible harvesting times.
- Shade Tolerance: Indicates whether the intercrop can tolerate shade, especially when intercropped with taller crops.
- Season: Ensures that both crops are suitable for the same planting season

Feature	**Description**
`crop-1`	The **main or primary crop** grown in the field.
`crop-2`	The **intercrop candidate** being considered to grow alongside `crop-1`.
`C-1 Water`	Water requirement level for the primary crop (`crop-1`), e.g., low/medium/high.
`C-2 Water`	Water requirement level for the intercrop (`crop-2`).
`C-1 Nutrients`	Nutrient requirement level for `crop-1`.
`C-2 Nutrients`	Nutrient requirement level for `crop-2`.
`C-1 root levels`	Root depth level of `crop-1` (e.g., shallow, medium, deep).
`C-2 root levels`	Root depth level of `crop-2`.
`Pest Disease-Overlap (%)`	Percentage of common pests/diseases between `crop-1` and `crop-2`.
`days-to-maturity diff`	Absolute difference in days to maturity between the two crops.
`shade tolerance`	Whether `crop-2` can tolerate shade, especially if `crop-1` has a tall canopy.
`Season`	The growing season both crops are suitable for (e.g., Kharif, Rabi, Zaid).

Table 4.2: Synthetic Dataset Features for Intercrop Recommendation

E) **Rule Based System:**

This stage does not involve any machine learning techniques but instead uses domain knowledge and predefined compatibility rules derived from agronomic practices. The rule for determining a *good match* between two crops is defined as:

```
good_match = 1 if (
    pest_overlap <= 30 and
    nutrient_diff >= 1 and
    c1_root != c2_root and
    days_to_maturity_diff in ['low', 'medium']
) else 0
```

Rule Descriptions

- **Pest Overlap $\leq$ 30%**: Ensures that the crops do not attract the same pests, reducing the risk of pest spread.
- **Nutrient Difference $\geq$ 1**: Promotes nutrient diversity. Nutrient levels are encoded as:

```
nutrient_levels = {'low': 1, 'medium': 2, 'high'
    : 3}
nutrient_diff = abs(nutrient_levels[c1_nutrients
    ] - nutrient_levels[c2_nutrients])
```

- **Different Root Structures**: The root types of the two crops should not be the same (`c1_root != c2_root`) to avoid competition and to use different soil depths.
- **Compatible Maturity Duration**: The difference in days to maturity should be classified as `'low'` or `'medium'`, allowing for compatible harvesting and resource use.

Crops that satisfy the conditions above are shortlisted as recommended intercrops.If multiple options are available, they can be ranked based on least pest overlap or highest maturity compatibility.

Chapter 5

Results

The classification models were evaluated based on their accuracy scores, as summarized in Table 5.1. Among the models tested, Random Forest achieved the highest accuracy at **99.3%**, closely followed by the Decision Tree model at **98.4%**, indicating the effectiveness of tree-based methods for the given dataset. Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) also performed well, with accuracies of **96.8%** and **95.6%**, respectively. Logistic Regression yielded a respectable accuracy of **96.3%**, whereas Linear Regression, which is not inherently designed for classification tasks, recorded a significantly lower accuracy of **9.31%**.

Model	**Accuracy**
`Linear Regression`	9.31%
`Logistic Regression`	96.3%
`Support Vector Machine`	96.8%
`K-Nearest Neighbors`	95.6%
`Decision Tree`	98.4%
`Random Forest`	99.3%

Table 5.1: performance metrics of step-1 using various models

These results underscore the superior performance of ensemble and non-linear models over linear models in handling the complexities of the dataset, making Random Forest and Decision Tree suitable candidates for further deployment or optimization

Following the evaluation and selection of the most accurate classification models (Random Forest and Decision Tree), the next step involved integrating a rule-based system to provide crop recommendations. This system uses two key inputs:

- **Predicted Crop (from the classification model output)**
- **Current Season**

The rule-based engine was designed using domain-specific agricultural knowledge, where specific crops are best suited for particular seasons due to factors such as temperature, rainfall, and soil conditions. Based on these rules, the system identifies and recommends the top two suitable crops for the given season and predicted output.

This hybrid approach, combining machine learning with expert-driven rule logic, improves the reliability and context-awareness of the recommendation system, making it more useful for real-world agricultural decision-making.

Chapter 6

Conclusion and Future work

This project successfully developed a hybrid system for crop recommendation by integrating machine learning classification models with a rule-based recommendation engine. Among the models evaluated, Random Forest and Decision Tree demonstrated the highest accuracy (99.3% and 98.4%, respectively), making them the most suitable for crop prediction tasks. The implementation of a rule-based system, informed by seasonal agricultural knowledge, further enhanced the system's effectiveness by recommending the intercrops possible given soil and environmental conditions and most appropriate for the current season.

To build upon the current work, several enhancements can be pursued:

- Incorporation of Real-Time Weather and Soil Data
- Expansion to Multi crop and Region-Specific Recommendations
- Use of Real time data for intercrops
- Use of Deep Learning Models

References

[1] M. R. Mahmood, M. A. Matin, S. K. Goudos, and G. Karagiannidis, "Machine learning for smart agriculture: A comprehensive survey," *IEEE Transactions on Artificial Intelligence*, 2023.

[2] Y. J. N. Kumar, V. Spandana, V. S. Vaishnavi, K. Neha, and V. G. R. R. Devi, "Supervised machine learning approach for crop yield prediction in agriculture sector," in *Proceedings of the Fifth International Conference on Communication and Electronics Systems (ICCES 2020)*, IEEE, 2020. IEEE Conference Record 48766.

[3] M. S. Rao, A. Singh, N. V. S. Reddy, and D. U. Acharya, "Crop prediction using machine learning," *Journal of Physics: Conference Series*, vol. 2161, p. 012033, 2022.

[4] E. Khosla, R. Dharavath, and R. Priya, "Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression," *Environment, Development and Sustainability*, vol. 22, pp. 5687–5708, Aug. 2020.

[5] H. Gebru, "A review on the comparative advantages of intercropping to mono-cropping system," *Journal of Biology, Agriculture and Healthcare*, vol. 5, no. 9, 2015.

[6] A. Ingle, "Crop recommendation dataset." https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset, 2020. Accessed: 2025-04-14.