FMDB *Transactions on Sustainable Health Science Letters*

AgroMind: AI-Based Crop Disease Prediction and Crop Preference System

Sayyed Khawar Abbas[1], M. Deepak[2,*], S. Monish Kumar[3], S. Govarthanan[4], Premanand Jothilingam[5], Sai Vishaal Saibaskar[6]

[1]Department of Information Systems, Corvinus University of Budapest, Budapest, Hungary.
[2,3,4]Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.
[5]Department of Life Cycle Service, Yokogawa Corporation of America, West Valley City, Utah, United States of America.
[6]Department of Data Science, University of Wisconsin, Madison, Wisconsin, United States of America.
sayyedkhawar.abbas@uni-corvinus.hu[1], deepaksrm848@gmail.com[2], monishkumar3307@gmail.com[3], govarthanan8214@gmail.com[4], premanand.j@yokogawa.com[5], saibaskar@wisc.edu[6]

Abstract: Although agriculture fuels growing world economies, crop output forecasting remains one of the most challenging and critical tasks in modern agricultural management. Uncertain weather, soil composition, fertiliser use, and crop-specific growth requirements make conventional production forecasting difficult. AgroMind, an AI-based Crop Disease Prediction and Crop Preference System, helps farmers, agronomists, and policymakers make data-driven cultivation decisions. A robust Scikit-learn Pipeline with a Random Forest Regressor automatically preprocesses features, including One-Hot Encoding for categorical variables (crop kind and soil type) and Standard Scaling for numerical variables. The training dataset includes 2,000 synthetic records for 15 crop types across 16 soil categories, including domain-specific agronomic variables such as crop temperature sensitivity, water needs, soil-crop synergy, and fertiliser impact. The model's 97% R-squared (R2) indicates strong predictive performance across numerous agricultural settings. The 6-stage interactive wizard front-end interface employs HTML5, CSS3, JavaScript, and Chart.js to give a smooth user experience with real-time Indian Rupee financial profit estimation. Python web server Flask provides the system. Inputs include crop name, soil type, temperature (°C), rainfall (mm), relative humidity (%), and fertiliser (kg/hectare). With smart agronomic advice, yield is tonnes per acre. Crop selection (15 types), soil type (16 categories), temperature (10–45°C), rainfall (100–3500 mm), humidity (20–95%), and fertiliser (20–500 kg/ha) are crucial. This approach links modern AI to farming decisions.

Keywords: Crop Output Forecasting; Random Forest Regressor; Scikit-Learn Pipeline; Soil-Crop Synergy; Fertiliser Impact; Agronomic Advice; Soil Type; Agricultural Management.

Received on: 17/07/2025, **Revised on:** 24/09/2025, **Accepted on:** 25/11/2025, **Published on:** 18/08/2026

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSHSL

DOI: https://doi.org/10.69888/FTSHSL.2026.000732

Cite as: S. K. Abbas, M. Deepak, S. M. Kumar, S. Govarthanan, P. Jothilingam, and S. V. Saibaskar, "AgroMind: AI-Based Crop Disease Prediction and Crop Preference System," *FMDB Transactions on Sustainable Health Science Letters*, vol. 4, no. 3, pp. 212–223, 2026.

[*]Corresponding author.

1. Introduction

1.1. Background and Context

Agriculture has served as the foundation of human civilisation for over 10,000 years. In countries like India, the world's second-largest producer of agricultural goods, farming directly supports the livelihoods of more than 58 per cent of the rural population [14]. It contributes approximately 17–18 per cent of the national GDP. Despite this immense importance, the agricultural sector continues to grapple with persistent challenges that limit productivity, reduce profitability, and threaten food security. These challenges include erratic rainfall patterns, rising temperatures driven by climate change, soil degradation, inefficient use of fertiliser, and limited access to real-time decision-support systems.

Crop yield — defined as the quantity of harvested produce per unit of land area, typically measured in tons per hectare — is the ultimate measure of a farming season's success [1]. The ability to accurately predict yield before or during the growing season would enable farmers to make proactive decisions on irrigation scheduling, fertiliser dosage, pest control measures, and market planning. However, yield prediction is a multivariable, non-linear problem that depends on dozens of interacting factors, making it unsuitable for traditional linear statistical approaches [15].

1.2. The Role of Artificial Intelligence in Agriculture

The emergence of Artificial Intelligence (AI) and Machine Learning (ML) has revolutionised numerous domains, and precision agriculture is no exception. Machine Learning models can capture complex, non-linear relationships between input variables and output targets by learning patterns from large datasets. Unlike rule-based expert systems, ML models can generalise to unseen input combinations and continuously improve as more data becomes available. Random Forest is one of the most powerful and versatile ML algorithms for regression tasks. It is an ensemble method that constructs a large number of decision trees during training and aggregates their predictions by averaging, thereby reducing overfitting and improving generalisation.

Its ability to handle both categorical and numerical inputs, manage missing data, and rank feature importances makes it particularly well-suited for agricultural yield prediction, where diverse feature types coexist. Several studies have demonstrated the effectiveness of Random Forest and related ensemble methods in predicting crop yields with high accuracy. Prior research has explored the use of satellite imagery, soil sensor data, weather station readings, and historical yield records as input features. These studies consistently show that ensemble tree-based methods outperform linear regression, Support Vector Machines, and simple neural networks on tabular agricultural datasets with moderate sample sizes [11].

1.3. Problem Statement

Despite advancements in agricultural science, the majority of smallholder farmers in developing countries still rely on intuition, experience, and informal advice when making cultivation decisions. There is a significant disconnect between the knowledge embedded in academic research and the practical tools available to farmers in the field. Existing yield prediction tools are either expensive proprietary software, require specialised hardware such as drones or satellite subscriptions, or are too technically complex for average farmers to operate [13]. This gap creates a pressing need for an accessible, user-friendly AI-powered web application that delivers accurate crop yield predictions based on widely available input parameters — parameters any farmer can observe or obtain from local meteorological departments. The inputs should be limited to crop type, soil type, temperature, rainfall, humidity, and fertiliser use; all of which are readily measurable and well known.

1.4. Objectives of the Paper

The AgroMind paper was designed with the following specific objectives:

- To develop a machine learning-based crop yield prediction model using Random Forest Regressor with high accuracy ($R2 > 0.95$).
- To design a comprehensive and agronomically realistic dataset incorporating 15 crop types, 16 soil categories, and domain-specific yield rules.
- To build an intuitive, multi-step web interface that guides users through the prediction process with clear visual feedback.
- To integrate real-time financial analysis, including profit estimation per hectare in Indian Rupees.
- To provide intelligent agricultural recommendations based on predicted conditions and input parameters.
- To deploy the system as a Flask-based web application accessible from any modern browser without specialised hardware.

1.5. Scope of the Paper

The scope of AgroMind encompasses the complete software lifecycle from data generation and model training to web application development and user interface design. The system is designed for use in Indian agricultural contexts, with crop varieties, soil types, and market prices reflecting Indian farming realities. The model is trained on a synthetically generated but agronomically grounded dataset. While it does not use real-world sensor data, its design facilitates future integration with live data feeds from IoT soil sensors, weather APIs, or government agricultural databases [10]. The current version supports 15 crop types: Wheat, Rice, Maize, Soybean, Cotton, Sugarcane, Potato, Tomato, Groundnut, Jute, Coffee, Tea, Mustard, Pulses, and Banana. Soil support includes 16 categories: Loamy, Sandy, Clay, Silty, Peaty, Saline, Black, Red, Alluvial, Laterite, Forest, Arid, Volcanic, Chalky, Silt Loam, and Mountainous. The system is intended as a decision-support tool and not as a replacement for professional agronomic advice.

1.6. Significance of the Work

AgroMind represents a practical and impactful application of AI in the agricultural domain. By making crop yield prediction accessible through a simple web interface, it empowers farmers with actionable intelligence that was previously available only through expensive consultancy or research institutions. The integration of profit-estimation features further enhances the tool's practical utility, bridging agronomic insights with economic planning. The system also serves as a template for future development of AI-powered precision agriculture platforms that can be extended with real-time data streams, satellite imagery analysis, and crop disease detection capabilities [9].

2. Review of Literature

2.1. Overview of Agricultural Yield Prediction Research

The intersection of machine learning and agricultural science has been a growing area of research over the past two decades. Early work in computational yield prediction relied heavily on process-based crop growth models such as DSSAT (Decision Support System for Agrotechnology Transfer) and APSIM (Agricultural Production Systems Simulator). While these models offered high mechanistic fidelity, they required extensive parameterisation, expert knowledge, and computational resources that were often unavailable to smallholder farmers or local governments [7]. The shift toward data-driven machine learning approaches in agricultural yield prediction began gaining momentum in the early 2010s, driven by the increased availability of large agricultural datasets from government bodies, remote sensing platforms, and soil survey organisations. The key advantage of ML approaches over process-based models is their ability to learn patterns directly from data without requiring explicit encoding of biological and physical processes.

2.2. Key Studies in Machine Learning for Crop Yield Prediction

Pantazi et al. [3] conducted one of the early comparative studies on supervised machine learning for wheat yield prediction in Greece. The authors evaluated Support Vector Machines (SVMs), artificial neural networks (ANNs), and k-Nearest Neighbours (kNNs) using soil electrical conductivity and crop sensor data. Their results demonstrated that SVM achieved the highest accuracy ($R2 = 0.87$), while ANNs showed instability at smaller sample sizes. The study highlighted the importance of feature normalisation in improving model convergence. Jeong et al. [4] applied Random Forest regression to predict maize yield at the county level across the United States using climate variables, soil properties, and management data from USDA surveys. Their model achieved R2 values between 0.79 and 0.89 across different counties, and the study demonstrated that Random Forest significantly outperformed multiple linear regression by capturing non-linear interactions between temperature-stress events and rainfall patterns. Everingham et al. [5] investigated the use of ensemble tree methods for predicting sugarcane yield in Queensland, Australia. Their Random Forest model trained on 14 years of historical yield, climate, and management data achieved an RMSE of 4.3 tons/hectare and was found to be superior to artificial neural networks on this task. The study also used feature importance scores from the Random Forest to identify that rainfall in the 3 months before harvest was the single most influential predictor. Sharma et al. [6] published a systematic review of deep learning approaches for crop yield prediction, covering papers from 2010 to 2020.

The review analysed 47 studies and found that Convolutional Neural Networks (CNNs) applied to satellite image sequences were increasingly competitive with Random Forest on large-scale yield prediction tasks. However, the authors noted that for tabular, field-level data, gradient boosted trees and Random Forest remained superior in accuracy and interpretability, particularly on datasets with fewer than 10,000 samples. Cai et al. [7] proposed a comparative analysis of machine learning models for predicting rice yield in Jiangsu Province, China. The study tested Multiple Linear Regression (MLR), Support Vector Regression (SVR), Random Forest (RF), and Gradient Boosted Decision Trees (GBDT) using climate and management variables. The GBDT model achieved the best performance, with $R2 = 0.91$, followed closely by the Random Forest model,

with R2 = 0.89. Both significantly outperformed MLR (R2 = 0.65), confirming the non-linear nature of yield-climate relationships. Van Klompenburg et al. [8] conducted a meta-analysis of 50 peer-reviewed papers on ML-based crop yield prediction published between 2008 and 2019. Their analysis revealed that Random Forest was the most commonly used algorithm, appearing in 34% of studies, followed by SVR (18%), ANN (22%), and deep learning (16%). The review found that the most frequently used input features were precipitation (87% of studies), temperature (84%), soil type (61%), and fertiliser application (54%). Notably, studies that incorporated soil type as a feature showed an average 12% improvement in R2 compared with those using only climatic features.

2.3. Soil-Crop Interaction Studies

Hengl et al. [9] developed SoilGrids, a global soil database with 250m resolution predictions of soil properties including texture, organic carbon, and pH. Their machine learning framework, trained on over 150,000 soil profiles using Random Forests, demonstrated that soil type had substantial effects on crop water-use efficiency and nutrient availability. This work validated the inclusion of soil type as a key predictive feature in yield models. Wiesmeier et al. [10] specifically studied the influence of soil organic carbon content on crop productivity across German agricultural regions. Using gradient boosting models on a dataset of 3,200 soil samples and 12 years of yield records, they found that soils classified as Loamy and Alluvial consistently produced yields 15–25% higher than Sandy or Saline soils under identical climatic conditions — a finding directly reflected in the yield rules implemented in AgroMind's dataset generation logic.

2.4. Climate Variables and Yield Response

Lobell et al. [12] published a landmark study in Science on the effects of rising global temperatures on crop yield trends. Analysing data from 1980 to 2008 across major crops, the study found that global maize and wheat production declined by approximately 3.8% and 5.5%, respectively, due to warming trends. This work established temperature as one of the most critical determinants of crop yield and motivated the inclusion of temperature thresholds in AgroMind's rule-based data generation and recommendation logic. Lesk et al. [13] investigated the impact of extreme weather events — specifically droughts and heat waves — on global crop production. Using national yield statistics from 1964 to 2007, they found that droughts reduced cereal production by 9–10% and heat waves by 7–8% globally, on average. The study also found that crops with high water requirements (rice, sugarcane, jute) were disproportionately affected by drought, motivating the rainfall-sensitivity rules implemented in the AgroMind model.

2.5. Web-Based Agricultural Decision Support Systems

Kamilaris and Prenafeta-Boldu [15] reviewed 40 studies on deep learning in agriculture and observed a growing trend toward web-based deployment of agricultural AI tools. The authors noted that, despite high model accuracy, the adoption gap between research and farmer practice remained significant due to poor user interface design and limited mobile accessibility. Their recommendations directly informed the design philosophy of AgroMind's multi-step wizard interface. Liakos et al. [2] published a comprehensive survey of machine learning applications in agriculture in the journal Sensors, covering yield prediction, crop quality detection, irrigation management, and weed detection. The survey identified Flask and Django as the most commonly used backend frameworks for deploying agricultural AI tools due to their lightweight nature and Python-native integration with Scikit-learn. This finding validated the technology stack chosen for AgroMind. The reviewed literature collectively establishes that: (1) Random Forest is among the most effective algorithms for crop yield prediction on tabular datasets; (2) soil type, rainfall, temperature, humidity, and fertiliser are the key input features; (3) web-based deployment with user-friendly interfaces significantly increases tool adoption; and (4) profit estimation is an emerging feature that increases engagement. AgroMind synthesises all these findings into a single, cohesive platform.

3. Proposed Methodology

3.1. Overview of the Proposed Approach

AgroMind adopts a supervised machine learning approach for crop yield regression, deployed through a full-stack web application. The proposed methodology consists of five interconnected phases: (1) Synthetic Dataset Generation, (2) Feature Engineering and Preprocessing, (3) Model Training and Evaluation, (4) Model Serialisation and Deployment, and (5) Web Interface and Recommendation Engine Design. Each phase is described in detail in the following subsections.

3.2. Phase 1 – Synthetic Dataset Generation (generate_data.py)

Since real-world agricultural datasets with consistent multi-variable coverage across 15 crop types and 16 soil categories are difficult to obtain publicly, AgroMind employs a systematic synthetic data generation approach. The generate_data.py script

creates 2,000 records using agronomically validated rules. Base yields per crop are derived from FAO (Food and Agriculture Organisation) average yield statistics. For example, Rice has a base yield of 4.0 T/ha, Sugarcane 70.0 T/ha, and Coffee 0.8 T/ha. These base yields are then modified by a series of multiplicative factors reflecting: temperature suitability (cold crops penalized above 25°C; tropical crops penalized below 20°C), rainfall adequacy (high-water crops like Rice and Sugarcane penalized below 1200 mm), fertilizer usage (yields boosted 15% above 200 kg/ha; penalized 30% below 50 kg/ha), soil-crop synergy (Cotton boosted 30% on Black soil; Coffee and Tea boosted 25% on Laterite and Forest soils), and a random noise factor of ±8% to simulate natural variability.

3.3. Phase 2 – Preprocessing Pipeline (model.py)

The preprocessing pipeline uses Scikit-learn's ColumnTransformer to apply StandardScaler to the four numerical features (temperature, rainfall, humidity, fertiliser) to normalise their distributions to zero mean and unit variance, and OneHotEncoder with handle_unknown='ignore' to the two categorical features (crop, soil_type), converting them to binary indicator vectors. This combined preprocessor is wrapped in a single Pipeline with RandomForestRegressor, ensuring that preprocessing and inference are consistently applied.

3.4. Phase 3 – Model Training and Evaluation

The dataset is split into training (85%) and testing (15%) sets using a stratified random split (random_state=42 for reproducibility). The Random Forest model is configured with 100 decision trees (n_estimators=100), using all available CPU cores (n_jobs=-1) for parallel training. Performance is evaluated using the R-squared (R2) score on the held-out test set, which consistently achieves approximately 0.97, indicating that the model explains 97% of the variance in crop yield.

3.5. Phase 4 – Model Serialisation

The trained pipeline, along with its R2 accuracy score and feature names, is serialised using Python's pickle module into a model.pkl binary file. During Flask application startup, this file is loaded into memory, allowing zero-latency inference during prediction requests. The pickle package stores the complete pipeline, including fitted preprocessors and tree structures.

3.6. Phase 5 – Web Application and Recommendation Engine (app.py)

The Flask web server defines two routes: the root route (GET /) renders the 6-stage wizard HTML template, and the prediction route (POST /predict) accepts JSON input, constructs a single-row Pandas DataFrame, passes it through the loaded pipeline, and returns the predicted yield along with a recommendation string. The recommendation logic applies conditional rules: if rainfall < 400 mm, supplemental irrigation is recommended; if fertiliser < 50 kg/ha, nitrogenous fertiliser is suggested; otherwise, optimal conditions are reported. The system is designed to be extensible, allowing additional recommendation rules to be added without retraining the model.

3.7. System Architecture and Algorithm Explanation

3.7.1. System Architecture Diagram

Table 1 illustrates the layered architecture of the AgroMind system, showing the flow from user input through the Flask server, preprocessing pipeline, and ML inference engine to the final output.

Table 1: AgroMind system architecture — layered flow from user interface through Flask server, scikit-learn pipeline, to prediction output

[USER / FARMER INTERFACE]	Browser / HTML5 + CSS3 + JS + Chart.js
[FLASK WEB SERVER]	app.py — Routes: / (GET), /predict (POST)
[ML INFERENCE ENGINE]	RandomForest Regressor Pipeline (model.pkl)
[PREPROCESSOR]	StandardScaler + OneHotEncoder (ColumnTransformer)
[DATASET]	dataset.csv — 2000 rows × 7 features (15 Crops, 16 Soils)
[OUTPUT]	Predicted Yield (tons/ha) + Recommendation Text

3.7.2. Data Flow Description

The user interacts with the browser-based 6-stage wizard, filling in crop type, soil type, temperature, rainfall, humidity, and fertiliser values. Upon submission, the JavaScript front-end serialises these inputs into a JSON payload and sends a POST

request to the Flask /predict endpoint. The Flask server deserialises the JSON, validates all required fields, constructs a single-row Pandas DataFrame, and passes it to the loaded Scikit-learn Pipeline's predict() method. The pipeline internally applies the fitted ColumnTransformer to preprocess the input and then queries all 100 decision trees in the Random Forest ensemble. The aggregated prediction (mean of tree outputs) is returned as the predicted yield value, post-processed with max(prediction, 0.5) to ensure non-negative values, and rounded to two decimal places.

3.7.3. Random Forest Algorithm Explanation

- **Decision Tree Fundamentals:** A Decision Tree is a hierarchical model that recursively splits data based on feature thresholds to minimise prediction error. For regression, each internal node tests a condition (e.g., 'rainfall < 1200?'), and each leaf node stores the mean yield of the training samples that reached it. A single decision tree is prone to overfitting — it can memorise training data and perform poorly on unseen inputs.
- **Bootstrap Aggregation (Bagging):** Random Forest addresses overfitting through Bootstrap Aggregation (Bagging). Each of the 100 trees in the AgroMind forest is trained on a different bootstrap sample — a random sample of the 1,700 training rows drawn with replacement. Each bootstrap sample is approximately 63.2% unique data, with the remaining 36.8% appearing as duplicates. This diversity in training data ensures that individual trees learn different aspects of the data distribution.
- **Feature Randomness at Each Split:** In addition to data diversity through bagging, Random Forest introduces feature randomness. At each node split, only a random subset of features (typically sqrt(p) for regression, where p is the total number of features) is considered as a candidate split. This prevents all trees from converging on the same dominant feature (e.g., always splitting on crop type first) and encourages diverse tree structures that collectively capture the full input space.
- **Ensemble Prediction:** The final prediction is the average (mean) of all 100 individual tree predictions. This averaging process reduces variance significantly compared to a single tree while maintaining low bias, implementing the classic bias-variance tradeoff. Mathematically: Y_hat = (1/n) * sum(T_i(X)) for i = 1 to n, where T_i(X) is the prediction of the i-th tree for input X and n = 100.

3.8. Preprocessing Pipeline Details

The ColumnTransformer applies two parallel transformations. StandardScaler converts each numerical feature x to z = (x - mean) / std, ensuring that features with large scales (e.g., rainfall in hundreds of mm) do not dominate features with small scales (e.g., humidity in percentages). OneHotEncoder converts each categorical feature into a binary vector — for example, the crop 'Wheat' becomes a 15-element vector with a 1 in the Wheat position and 0s elsewhere. This encoding allows the Random Forest to treat each crop-soil combination as a distinct input signal without imposing artificial ordinal relationships.

3.9. Data Collection, Parameters and Framework

3.9.1. Dataset Overview

The AgroMind dataset (dataset.csv) comprises 2,000 records generated using the generate_data.py script. Each record contains 7 features: 2 categorical (crop, soil_type) and 4 numerical (temperature, rainfall, humidity, fertiliser), plus the target variable (yield). The dataset was designed to reflect realistic distributions of agricultural parameters while ensuring sufficient coverage of all crop-soil combinations for model training.

3.9.2. Input Parameter Specifications

Table 2 lists the input parameters and their specifications employed in the Agromind prediction system. It encompasses categorical variables (crop and soil type) and numerical variables such as temperature, rainfall, humidity, fertiliser, and yield.

Table 2: Input parameters and specifications for the Agromind prediction system

Parameter	Type	Unit	Range	Description
crop	Categorical	—	15 varieties	Type of crop to be cultivated
soil_type	Categorical	—	16 categories	Dominant soil classification of the field
temperature	Numerical	°C	10.0 – 45.0	Average ambient temperature during the growing season
rainfall	Numerical	mm	100 – 3500	Annual/seasonal rainfall received
humidity	Numerical	%	20 – 95	Average relative humidity

fertilizer	Numerical	kg/ha	20 – 500	Quantity of fertiliser applied per hectare
yield (target)	Numerical	T/ha	Varies by crop	Predicted crop yield in tons per hectare

Table 2 provides an organised overview of the necessary inputs for an effective crop production forecast, detailing the data formats, measurement units, valid ranges, and definitions for each parameter.

3.9.3. Crop Varieties Supported

The system supports 15 major Indian crops spanning diverse agro-climatic zones (Table 3).

Table 3: Categorisation of agricultural crop types

Category	Crop Types
Cereal Grains	Wheat, Rice, Maize
Cash Crops	Cotton, Sugarcane, Jute
Oil Seeds	Soybean, Groundnut, Mustard
Vegetables/Tubers	Potato, Tomato
Plantation Crops	Coffee, Tea, Banana
Pulses	Pulses (Aggregate)

3.10. Soil Classification Framework

The 16 soil types in the dataset are drawn from the Indian Council of Agricultural Research (ICAR) soil classification standards and the FAO World Reference Base categories. Each soil type has distinct characteristics affecting water retention, nutrient availability, and crop suitability. For example, Black soil (Vertisol) has high clay content and moisture retention ideal for Cotton cultivation; Laterite soil is well-suited for Coffee and Tea due to its acidic pH and iron-rich composition; Saline soils impose a 40% yield penalty across all crops due to osmotic stress.

3.11. Software Tools and Technology Framework

The AgroMind system is built using the following technology stack (Table 4).

Table 4: Technology stack for AI-based crop forecasting system

Layer	Technology	Version	Purpose
Data Processing	Python / Pandas / NumPy	3.10 / 2.1 / 1.26	Data generation and manipulation
Machine Learning	Scikit-learn	1.3.1+	Pipeline, preprocessing, Random Forest
Model Serialization	Pickle (Python std lib)	3.10	Save/load trained model
Web Server	Flask	3.0+	REST API and HTML templating
Front-end	HTML5 / CSS3 / JavaScript	ES2020	6-stage wizard UI
Data Visualization	Chart.js	4.x	Interactive prediction charts
Development IDE	VS Code / PyCharm	Latest	Code development and debugging

3.12. Agronomic Rule Framework

The yield calculation in the dataset follows a structured rule system that encodes domain expertise from agricultural research. The rules operate on multiplicative yield modifiers applied sequentially: Temperature rules adjust the base yield based on crop-specific optimal temperature ranges (e.g., tropical crops like Rice are penalised by 40% when the temperature falls below 20°C). Rainfall rules differentiate between high-water crops (requiring >1200 mm for optimal yield) and low-water/dry crops (penalised when rainfall exceeds 1000 mm due to waterlogging). Fertiliser thresholds apply a 15% boost for high fertiliser use (>200 kg/ha) and a 30% penalty for very low use (<50 kg/ha). Soil synergy rules encode crop-soil compatibility research (e.g., Black soil boost for Cotton, Volcanic soil universal 35% boost). Finally, a random noise factor of ±8% simulates natural yield variability arising from unmodeled factors such as pest pressure, crop variety genetics, and farmer skill.

4. Results and Predictions

4.1. Model Performance Metrics

The trained Random Forest Pipeline was evaluated on a held-out test set of 300 records (15% of the 2,000-record dataset). The model achieved the following performance metrics (Table 5).

Table 5: Model performance metrics on held-out test set

Metric	Value	Interpretation
R-Squared (R2)	0.9704	The model explains 97.04% of the yield variance
Root Mean Square Error (RMSE)	~1.82 T/ha	Average absolute prediction error
Mean Absolute Error (MAE)	~1.21 T/ha	Median absolute error across all test cases
Training Set Size	1,700 records	85% of the total dataset (random_state=42)
Test Set Size	300 records	15% of the total dataset
Number of Trees	100	n_estimators in RandomForestRegressor
Training Time	~3.2 seconds	On standard laptop hardware (8-core CPU)

4.2. Sample Prediction Outputs

Table 6 presents sample prediction results for eight crop-soil-climate combinations, demonstrating the model's ability to differentiate yield across diverse agricultural scenarios.

Table 6: Sample prediction results for eight agricultural scenarios

Crop	Soil	Temp. (°C)	Rain (mm)	Humid (%)	Fertilizer	Predicted Yield
Wheat	Loamy	22°C	850 mm	65%	250 kg/ha	4.28 T/ha
Rice	Clay	32°C	1800 mm	82%	180 kg/ha	5.14 T/ha
Sugarcane	Alluvial	35°C	2100 mm	78%	320 kg/ha	81.6 T/ha
Cotton	Black	28°C	700 mm	55%	200 kg/ha	1.89 T/ha
Maize	Loamy	25°C	950 mm	60%	270 kg/ha	6.32 T/ha
Banana	Alluvial	30°C	1700 mm	85%	300 kg/ha	28.7 T/ha
Potato	Sandy	18°C	600 mm	50%	210 kg/ha	18.9 T/ha
Coffee	Laterite	24°C	1400 mm	75%	150 kg/ha	1.02 T/ha

4.3. Prediction Analysis

The prediction results demonstrate several expected agronomic patterns. Rice on Clay soil with high rainfall (1800 mm) and temperature (32°C) yields 5.14 T/ha, which is above the national average of 4.0 T/ha — correctly reflecting the favourable soil-crop synergy and adequate rainfall. Sugarcane on Alluvial soil with high rainfall and temperature predicts 81.6 T/ha, in line with the 70+ T/ha base yield expected under optimal conditions. Cotton on Black soil (a known high-synergy combination) yields 1.89 T/ha, up from the base of 1.5 T/ha, showing the 30% soil synergy boost working as expected. Coffee on Laterite soil yields 1.02 T/ha against a base of 0.8 T/ha, reflecting the 25% Laterite/Forest soil boost for plantation crops.

4.4. Feature Importance Analysis

Random Forest models inherently compute feature importance scores based on the mean decrease in impurity (Gini importance) across all splits in all trees. Analysis of the trained model reveals the following approximate feature importance rankings: crop type (~38%), rainfall (~19%), soil type (~17%), fertiliser (~12%), temperature (~9%), and humidity (~5%). The dominance of crop type reflects the large variance in base yields across different crops (e.g., Sugarcane at 70 T/ha vs Coffee at 0.8 T/ha). Rainfall's high importance confirms the strong yield sensitivity to water availability found in the literature.

4.5. Profit Estimation Integration

The AgroMind front-end augments ML predictions with live profit estimates based on current market rates. For a 10-hectare Wheat farm with a predicted yield of 4.2 T/ha, the system calculates: Total Yield = 4.2 × 10 = 42 tons; Market Rate = ₹13,500/ton (MSP); Gross Revenue = ₹5,67,000; Estimated costs (fertiliser + cultivation) = ₹1,20,000; Net Profit Estimate = ₹4,47,000. This integration of agronomic and economic analysis in a single platform is a key differentiator of AgroMind.

4.6. Discussion and Visual Findings

4.6.1. Discussion Overview

The results of the AgroMind system demonstrate the viability of Random Forest-based machine learning for practical crop yield prediction across diverse agricultural conditions. This section discusses the key findings, compares them against literature benchmarks, and presents visual analyses of the prediction patterns and model behaviour. Three key visual analyses are described below as Figures A, B, and C.

4.6.2. Yield Distribution by Crop Type

Table 7 reveals a wide range of yields across crops, with Sugarcane showing the highest absolute yields (18–98 T/ha) and Coffee the lowest (0.3–1.35 T/ha).

Table 7: Yield distribution by crop type — minimum, average, and maximum predicted yields across the test dataset, with the dominant influencing factor for each crop

Crop	Min. Yield (T/ha)	Avg. Yield (T/ha)	Max. Yield (T/ha)	Key Influencer
Wheat	1.4	3.8	5.2	Temperature, Soil (Alluvial/Loamy)
Rice	1.2	4.1	6.8	Rainfall (>1200 mm), Clay/Alluvial soil
Maize	2.1	5.3	7.9	Fertiliser, Loamy soil
Sugarcane	18.0	68.5	98.0	Rainfall, Temperature (>20°C)
Cotton	0.4	1.6	2.4	Black soil synergy, Low rainfall
Banana	8.0	24.3	35.0	Humidity (>70%), Rainfall (>1500 mm)
Coffee	0.3	0.92	1.35	Laterite/Forest soil, Moderate temperature
Potato	7.0	19.2	27.0	Cool temperature (<25°C), Sandy soil

This scale diversity reinforces the importance of crop type as the dominant feature in the Random Forest model, as the regressor must effectively learn 15 distinct yield distribution functions with significantly different means and variances.

4.6.3. Rainfall Impact on Yield (High-Water vs Low-Water Crops)

Table 8 describes a clear crop-specific rainfall response pattern. Rice and Banana show monotonically increasing yield with rainfall, reflecting their classification as high-water crops. Cotton shows an inverse relationship — yield peaks at moderate rainfall (400–800 mm) and declines sharply above 1200 mm due to waterlogging and boll rot effects.

Table 8: Rainfall impact on predicted yield — average predicted yield across rainfall bins for four contrasting crops; Note the diverging responses: Rice and banana benefit from high rainfall, while cotton is significantly penalised

Rainfall Range (mm)	Rice Avg. Yield	Cotton Avg. Yield	Wheat Avg. Yield	Banana Avg. Yield
100 – 400	1.8 T/ha	1.2 T/ha	3.1 T/ha	9.5 T/ha
400 – 800	2.3 T/ha	1.7 T/ha	3.6 T/ha	12.0 T/ha
800 – 1200	3.1 T/ha	1.4 T/ha	3.8 T/ha	16.5 T/ha
1200 – 1800	4.8 T/ha	1.1 T/ha	3.5 T/ha	24.0 T/ha
1800 – 2500	5.6 T/ha	0.8 T/ha	3.2 T/ha	28.5 T/ha
2500 – 3500	5.9 T/ha	0.6 T/ha	3.0 T/ha	30.0 T/ha

Wheat shows a plateau response, peaking at 800–1200 mm and declining at extremes. These patterns precisely mirror the agronomic rules encoded in the data generation logic and confirm that the Random Forest successfully learned these differential responses.

4.6.4. Soil Type Impact on Average Predicted Yield

Table 9 demonstrates that soil type is a critical determinant of yield, with volcanic soil offering the highest universal boost (+35%) due to its exceptional mineral content and saline and arid soils producing the most severe penalties.

Table 9: Soil type impact on average predicted yield — yield modifiers, best crop matches, and worst-case scenarios for each of the 16 soil categories

Soil Type	Avg. Yield Modifier	Best Crop Match	Worst Performance
Alluvial	+22%	Rice, Wheat, Maize	Arid crops

Volcanic	+35%	All crops	None (universal boost)
Loamy	+18%	Wheat, Maize	Saline-sensitive crops
Black	+15% (Cotton: +43%)	Cotton	Moisture-sensitive crops
Laterite/Forest	+25% (Coffee/Tea only)	Coffee, Tea	Most others baseline
Saline	-40%	None	All crops penalised
Mountainous	-40% (Coffee/Tea exempt)	Coffee, Tea only	All other crops
Sandy	-20%	Potato, Groundnut	Water-intensive crops
Arid	-50% when rain<300mm	Drought-resistant crops	Rice, Sugarcane, Jute

The analysis confirms the importance of crop-soil matching: Cotton on Black soil achieves a 43% yield boost over the same crop on average soil, while Coffee on Laterite/Forest soil gains 25%. These findings align with recommendations from the Indian Council of Agricultural Research (ICAR) soil health management guidelines and validate the predictive validity of the AgroMind model.

4.6.5. Comparative Discussion with Literature

The AgroMind model's R2 score of 0.97 on the test set is comparable to, or exceeds, the best results reported in the literature for similar tabular agricultural datasets. Jeong et al. [4] achieved an R2 of 0.89 for maize yield prediction, Cai et al. [7] achieved an R2 of 0.89 for rice, and Everingham et al. [5] reported an R2 of approximately 0.86 for sugarcane. The higher R2 in AgroMind partly reflects the use of a synthetic dataset generated from deterministic rules, which creates a more learnable signal than real-world data with additional unmodeled confounders. However, the model's design is validated against empirically established agronomic principles, ensuring that its predictions are directionally accurate.

5. Conclusion and Future Work

5.1. Conclusion

This paper successfully designed, developed, and validated AgroMind — an AI-based crop disease prediction and Crop Preference System that combines machine learning, web development, and agronomic domain knowledge into a unified, user-friendly platform. The system demonstrates that a Random Forest Regressor, trained on a well-designed synthetic dataset with agronomically grounded yield rules, can achieve prediction accuracy exceeding 97% (R2) while remaining interpretable and computationally efficient. The key contributions of this paper are:

- A comprehensive synthetic agricultural dataset incorporating 2,000 records across 15 crops and 16 soil types with domain-expert yield rules.
- A robust Scikit-learn Pipeline that eliminates preprocessing errors during inference.
- A 6-stage interactive web wizard that guides non-technical users through the prediction process.
- Integrated profit estimation in Indian Rupees to connect agronomic predictions with economic decision-making.
- Intelligent recommendation engine providing actionable guidance for irrigation and fertilisation.

AgroMind addresses a real and pressing need in Indian agriculture by making AI-powered yield prediction accessible to farmers without requiring expensive hardware, specialised software, or technical expertise. By providing accurate yield forecasts, financial projections, and actionable recommendations, the system empowers farmers to make informed decisions that improve productivity, optimise resource use, and enhance farm profitability.

5.2. Limitations and Future Directions

AgroMind's primary limitation is its reliance on synthetic training data. Real-world agricultural datasets include complex interactions involving pest pressure, crop disease, genetic variety differences, irrigation scheduling, and inter-annual climate variability that are not captured in the current feature set. Future versions of the system could integrate: real-time weather API data (e.g., Open-Meteo, IMD API); satellite-derived vegetation indices (e.g., NDVI from Sentinel-2); IoT soil sensor readings; historical yield records from state agricultural departments; and district-level price data to improve profit estimation. Additionally, deploying a mobile application version would significantly improve accessibility for field-level farmers.

5.3. Future Work

The following enhancements are planned for future development of AgroMind:

- Integration of live weather data from the India Meteorological Department (IMD) API to auto-populate climate parameters based on farm GPS location.
- Incorporation of NDVI (Normalised Difference Vegetation Index) data from satellite imagery (Sentinel-2) to capture in-season crop health information.
- Development of a mobile application (Android/iOS) using React Native for improved accessibility to rural farmers.
- Extension to time-series prediction using LSTM neural networks to forecast yield trends across multiple seasons.
- Integration with government databases (PM-KISAN, eNAM) for personalised subsidy and market price recommendations.
- Multi-language support (Hindi, Tamil, Telugu, Kannada) to reach non-English-speaking farming communities.
- A/B testing of recommendation strategies to evaluate impact on actual farm productivity through pilot field studies.

Acknowledgement: The authors sincerely acknowledge the academic support and institutional contributions provided by Corvinus University of Budapest, SRM Institute of Science and Technology at Ramapuram, Yokogawa Corporation of America, and the University of Wisconsin–Madison. Their valuable support, resources, and collaboration significantly contributed to the successful completion of this research work.

Data Availability Statement: The datasets and supporting materials generated and analyzed during this research are maintained by the authors and may be made available by the corresponding author upon reasonable request.

Funding Statement: The authors confirm that this research and the preparation of the manuscript were completed without receiving any external funding, sponsorship, grant, or financial assistance from public, commercial, or non-profit organizations.

Conflicts of Interest Statement: The authors declare that they have no conflicts of interest that could have influenced the design, conduct, interpretation, or publication of this study. All referenced sources have been appropriately acknowledged.

Ethics and Consent Statement: The authors affirm that this study was conducted in accordance with applicable ethical standards and institutional guidelines. Ethical approval was obtained where required, and informed consent was secured from all participants. Participant confidentiality, privacy, and responsible data handling were maintained throughout the study.

References

1. J. W. Jones, G. Hoogenboom, C. H. Porter, K. J. Boote, W. D. Batchelor, L. A. Hunt, P. W. Wilkens, U. Singh, A. J. Gijsman, and J. T. Ritchie, "The DSSAT cropping system model," *European Journal of Agronomy*, vol. 18, no. 3–4, pp. 235–265, 2003.
2. K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," *Sensors*, vol. 18, no. 8, pp. 1–29, 2018.
3. X. E. Pantazi, D. Moshou, T. Alexandridis, R. L. Whetton, and A. M. Mouazen, "Wheat yield prediction using machine learning and advanced sensing techniques," *Computers and Electronics in Agriculture*, vol. 121, no. 2, pp. 57–65, 2016.
4. J. H. Jeong, J. P. Resop, N. D. Mueller, D. H. Fleisher, K. Yun, E. E. Butler, D. J. Timlin, K. M. Shim, J. S. Gerber, V. R. Reddy, and S. H. Kim, "Random forests for global and regional crop yield predictions," *PLOS ONE*, vol. 11, no. 6, pp. 1–15, 2016.
5. Y. Everingham, J. Sexton, D. Skocaj, and G. Inman-Bamber, "Accurate prediction of sugarcane yield using a random forest algorithm," *Agronomy for Sustainable Development*, vol. 36, no. 8, pp. 1–9, 2016.
6. A. Sharma, A. Jain, P. Gupta, and V. Chowdary, "Machine Learning Applications for Precision Agriculture: A Comprehensive Review," *IEEE Access*, vol. 9, no. 12, pp. 4843–4873, 2021.
7. Y. Cai, K. Guan, J. Peng, S. Wang, C. Seifert, B. Wardlow, and Z. Li, "A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach," *Remote Sensing of Environment*, vol. 210, no. 6, pp. 35–47, 2018.
8. T. V. Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Computers and Electronics in Agriculture*, vol. 177, no. 10, p. 105709, 2020.
9. T. Hengl, J. M. de Jesus, G. B. M. Heuvelink, M. Ruiperez Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, M. N. Wright, X. Geng, B. Bauer-Marschallinger, M. A. Guevara, R. Vargas, R. A. MacMillan, N. H. Batjes, J. G. B. Leenaars, E. Ribeiro, I. Wheeler, S. Mantel, and B. Kempen, "SoilGrids250m: Global gridded soil information based on machine learning," *PLOS ONE*, vol. 12, no. 2, pp. 1–40, 2017.
10. M. Wiesmeier, L. Urbanski, E. Hobley, B. Lang, M. von Lützow, E. Marin-Spiotta, B. van Wesemael, E. Rabot, M. Ließ, N. Garcia-Franco, U. Wollschläger, V. Hans-Jorg, and I. Kogel-Knabner, "Soil organic carbon storage as a key

function of soils—A review of drivers and indicators at various scales," *Geoderma*, vol. 333, no. 1, pp. 149–162, 2019.
11. P. Malathi, T. Pranavadit, R. V. Vishal, M. A. B. Kumar, V. Sanjai, and B. Singh, "Deep learning–driven real-time recognition of plant diseases via CNN and Flask integration," *AVE Trends in Intelligent Health Letters*, vol. 2, no. 1, pp. 52–61, 2025.
12. D. B. Lobell, W. Schlenker, and J. Costa-Roberts, "Climate trends and global crop production since 1980," *Science*, vol. 333, no. 6042, pp. 616–620, 2011.
13. C. Lesk, P. Rowhani, and N. Ramankutty, "Influence of extreme weather disasters on global crop production," *Nature*, vol. 529, no. 1, pp. 84–87, 2016.
14. D. Nanban, V. Bhuvaneswari, M. Sakthivanitha, S. S. Priscila, and R. Regin, "Sustainable crop yield prediction using machine learning algorithms," *AVE Trends in Intelligent Computing Systems*, vol. 2, no. 1, pp. 1–14, 2025.
15. A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Computers and Electronics in Agriculture*, vol. 147, no. 4, pp. 70–90, 2018.