📄 SimAgro: AI-Powered Simulation for Real-Time Crop Monitoring and Yield Prediction

Abstract

This paper presents **SimAgro**, a smart agricultural simulation system that leverages **Machine Learning** to predict **crop health** and **yield** using dynamic and static environmental and agronomic parameters. The project integrates over **30 features**, including vegetation indices, weather data, pest metrics, and soil properties, collected across multiple CSV datasets. The system achieves high accuracy (98.84%) in crop health classification and a strong R^2 score (0.9719) in yield prediction, making it viable for real-time deployment in smart farming ecosystems.

Keywords

Precision Agriculture, Crop Health, Yield Prediction, Machine Learning, Random Forest, Smart Farming, Streamlit Dashboard, AI in Agriculture, Agriculture 6.0

I. Introduction

Agriculture remains a foundational pillar of global economies, yet is increasingly vulnerable to climate change, pest outbreaks, and resource mismanagement. Traditional methods of yield estimation and crop health monitoring are both **labor-intensive and inconsistent**. To address this, **SimAgro** provides a **simulation-based solution** that predicts real-time crop conditions using AI models trained on **sensor-inspired features** and **environmental datasets**.

II. Literature Review

Existing works in crop prediction rely on either remote sensing data or basic statistical models. However, few integrate both static (soil type, crop type) and dynamic (weather, NDVI, rainfall)

features into a cohesive ML framework. Moreover, many models lack scalability and adaptability. SimAgro bridges this gap by utilizing **Random Forest-based ensemble learning** and a **Streamlit-powered dashboard** for simulation and decision support.

III. System Architecture

A. Input Data

- **Sources**: Curated CSVs & XLSX files combining static crop features and dynamic seasonal parameters.
- **Total Features**: 30 (manually selected from 95 possible inputs).

Selected Features Include:

- NDVI, SAVI, Leaf Area Index, Soil Moisture, Fertilizer Used
- Temperature, Rainfall, Humidity, Pest Level, Crop Height
- Irrigation Used, Days to Harvest, Sunlight Hours, Soil pH, etc.

B. Pipeline Workflow

1. **Data Cleaning & Merging**: Datasets from `main_features.xlsx` and `dynamic_features.xlsx` are preprocessed.
2. **Label Encoding**: Applied to categorical columns (e.g., Crop, Soil Type).
3. **Health Classification**: 3-class problem based on health score (Poor, Average, Good).
4. **Yield Regression**: Predicting tons/hectare using Random Forest.
5. **Model Saving**: Trained models stored via `joblib`.
6. **Visualization**: Dashboard built using **Streamlit**.

IV. Model Implementation

A. Health Prediction Model

- **Algorithm**: Random Forest Classifier
- **Classes**:
  - Poor: Health Score < 40
  - Average: 40 ≤ Score < 70
  - Good: Score ≥ 70
- **Accuracy**: 98.84%
- **Classification Report**:

- Precision: 0.99–1.00
- Recall: 0.80–1.00
- F1-Score: Weighted Avg = 0.99

Confusion Matrix indicates excellent generalization across categories.

B. Yield Estimation Model

- **Algorithm**: Random Forest Regressor
- **Target**: `Yield_tons_per_hectare`
- **R^2 Score**: 0.9719
- **RMSE**: 0.2844 tons/hectare

These metrics suggest high prediction accuracy even under multi-feature variability.

V. Dashboard UI (SimAgro Streamlit)

- Real-time sliders to simulate:
  - Temperature, Soil Moisture, Pest Level, Fertilizer, Crop Height, etc.
- Cobot Advisor: Provides smart suggestions for irrigation, pest control, and fertilizer adjustment.
- Graphs for:
  - Crop growth trends
  - Health and yield estimates
  - Past simulation results
- Dynamic visualization of:
  - Confusion Matrix
  - Yield RMSE Distribution

VI. Results and Discussion

SimAgro successfully merges **data simulation** with **predictive modeling** for both health and yield. The health model's near-perfect classification accuracy indicates strong learning across diverse crop conditions, while the yield model's low RMSE highlights excellent real-world applicability. Minor variance in the "Average" health class recall suggests potential for further improvement through class balancing or ensemble hybridization.

VII. Tools and Technologies Used

Component	Tool / Framework
Programming Language	Python
ML Libraries	Scikit-learn, Pandas, NumPy
Visualization	Matplotlib, Seaborn
Web UI	Streamlit
Data Format	CSV, XLSX
Model Storage	Joblib
IDE	VS Code

VIII. Conclusion

SimAgro provides a scalable and intelligent framework for crop monitoring and yield prediction. By incorporating a variety of static and dynamic features, it offers **accurate, real-time advisory support** to farmers and agri-scientists alike. Its modular and visual nature also makes it a powerful educational tool.

IX. Future Scope

- Add LSTM-based models for sequential predictions
- Introduce historical satellite image integration
- Extend to more crop types and regional conditions
- Integrate into IoT-based smart farming kits
