📄 SimAgro: A Hybrid AI Framework for Crop Health, Yield, and Disease Prediction

Abstract

Agriculture is undergoing a digital transformation, with Artificial Intelligence (AI) playing a pivotal role in precision farming. This paper proposes **SimAgro**, an integrated machine learning and deep learning framework for **crop health monitoring, yield estimation, and disease detection**. SimAgro leverages **feature-based models** trained on agronomic and environmental data alongside **image-based convolutional neural networks (CNNs)** for disease identification.

Experiments demonstrate:

- **98.84% accuracy** in crop health classification using Random Forest.
- **R^2 score of 0.9719** for yield estimation, with RMSE of 0.28 tons/hectare.
- **98% accuracy** in leaf disease detection using ResNet18 trained on the **PlantVillage dataset**.

The proposed system is deployed via an interactive **Streamlit dashboard**, enabling real-time farmer advisory support through data-driven predictions and actionable recommendations.

Keywords: Precision Agriculture, Crop Health, Yield Prediction, Leaf Disease Detection, Machine Learning, Deep Learning, ResNet18, PlantVillage Dataset, Smart Farming.

I. Introduction

Agriculture accounts for a significant share of global GDP and food supply, yet it faces challenges from **climate variability, pest outbreaks, and plant diseases**. Traditional monitoring techniques rely on **manual inspection**, which is time-consuming, error-prone, and limited in scalability.

Recent works have applied **machine learning** for yield prediction and **deep learning** for disease detection, but few offer an **end-to-end integrated system** covering **health, yield, and disease diagnostics** in a single pipeline.

SimAgro bridges this gap by combining:

1. **Health classification** based on multi-feature agronomic data.
2. **Yield estimation** through regression models.
3. **Leaf disease detection** using CNNs trained on PlantVillage.
4. **Streamlit-based decision support system** with a virtual cobot advisor.

II. Related Work

- **Yield Prediction**: Studies often rely on regression models (e.g., linear regression, SVMs) using rainfall, temperature, and soil data. These models lack adaptability to dynamic environments.
- **Disease Detection**: Deep learning approaches (CNN, VGG16, ResNet) on PlantVillage dataset have achieved >95% accuracy. However, most are standalone applications without integration into broader farming systems.
- **Integrated Frameworks**: Very few works combine **feature-based crop modeling** and **image-based disease detection** into one platform, highlighting a research gap.

III. Methodology

A. Data Sources

- **Tabular Data**: Custom agronomic datasets (30+ features) including NDVI, SAVI, soil moisture, pest level, crop height, fertilizer use, rainfall, temperature, sunlight, and soil pH.
- **Image Data**: PlantVillage dataset from Kaggle, containing **54,000+ labeled leaf images across 38 crop-disease classes**.

B. Preprocessing

- Feature scaling and label encoding for categorical features (crop type, soil, region).
- Image augmentation (rotation, flipping, brightness variation) to improve CNN robustness.

C. Models

1. **Health Classification**
   - Algorithm: Random Forest Classifier
   - Accuracy: 98.84%
2. **Yield Estimation**
   - Algorithm: Random Forest Regressor
   - $R^2 = 0.9719$, RMSE = 0.28
3. **Disease Detection**
   - Model: ResNet18 CNN (pretrained on ImageNet, fine-tuned on PlantVillage).
   - Accuracy: 98%
   - Provides disease class, cause, and treatment recommendation via a knowledge base.

D. Deployment Architecture

- Models serialized with `joblib` and PyTorch.
- Integrated into a **Streamlit dashboard**:
  - **Sliders** for input features.
  - **Image uploader** for disease detection.
  - **Virtual Cobot Advisor** for recommendations.
  - **Charts** for visualization of simulation trends.

IV. Results

Task	Algorithm	Dataset	Accuracy / R^2	Remarks
Crop Health	Random Forest Classifier	Agronomic dataset	**98.84%**	Robust across Poor, Average, Good
Yield Prediction	Random Forest Regressor	Agronomic dataset	**R^2 = 0.9719**, RMSE=0.28	Low error, scalable
Disease Detection	ResNet18 CNN	PlantVillage (Kaggle)	**98%**	Strong generalization across 38 classes

V. Comparative Analysis of Existing Projects

Task	Algorithm / Model	Dataset	Accuracy / R^2	Remarks
Crop Disease Detection (IoT Pivot System)	ResNet50 (embedded)	Custom field images	**99.8%**	Real-time pivot system sprays automatically; hardware-dependent, less scalable.
Multi-crop Disease Detection (Slender-CNN)	Lightweight CNN	CRW (Corn, Rice, Wheat) dataset	**88.54%**	Efficient but lower accuracy; crop-specific.
Agro Deep Learning Framework (ADLF)	Deep Learning (Hybrid)	Environmental data (soil, temp, humidity)	**85% (F1=88.9%)**	Focused on environmental features, no disease detection.
SMARD	CNN	Tomato leaf images	**97.3% (F1=96%)**	Focused on tomato, not multi-crop; high accuracy.

Task	Algorithm / Model	Dataset	Accuracy / R²	Remarks
FourCropNet	CNN + Attention	Grape, corn, soybean, cotton	**95–99.7%**	Strong for specific crops, but not yield-focused.
sCrop IoT Device	CNN (on-device)	Real-field images	**99.2%**	Solar-powered IoT; hardware cost limits adoption.
Plantix (Mobile App)	CNN	Farmer-uploaded images	~**95%**	Widely used, but lacks yield prediction & feature integration.
Aggrotech (Tomato)	VGG19, Inception v3	Tomato leaf dataset	**93.9%**	Tomato-only, lower generalization.
Smart Agriculture Competition (AI Greenhouse)	ML + IoT	Real greenhouse strawberry data	Yield ↑196%, Cost ↓75.5%	Amazing results, but limited to controlled greenhouse setting.
IIT KGP Robot	CV + ML	Pest detection dataset (field)	Not disclosed	Strong robotics integration, but heavy hardware.

VI. Why SimAgro is Better

SimAgro outperforms existing systems due to its **hybrid and integrated approach**:

1. Combines **three tasks in one system**: crop health classification, yield prediction, and leaf disease detection.
2. Achieves **high accuracy across all tasks**: 98.84% (health), R^2=0.9719 (yield), 98% (disease).
3. Accepts **multi-modal inputs**: tabular agronomic features + image inputs.
4. Farmer-friendly deployment via **Streamlit dashboard** with cobot advisor, sliders, and charts.
5. **Hardware-independent** and scalable for smallholder farmers (unlike IoT pivots or robots).
6. Provides **knowledge-based disease insights** (cause + treatment) alongside predictions.

Thus, **SimAgro is not just accurate but also practical, scalable, and holistic**, addressing multiple aspects of smart farming in one solution.

VII. Discussion

- The **Random Forest models** demonstrated excellent stability across diverse feature sets, suitable for real-world deployment.
- The **ResNet18 disease detector** achieved high accuracy, validating CNNs for agricultural imaging tasks.
- The **integration into a single dashboard** enhances usability for farmers, researchers, and policymakers.
- Comparative analysis shows that while other projects excel in **specialized areas**, SimAgro is unique in combining multiple predictive dimensions into one platform.

VIII. Conclusion

SimAgro presents a **hybrid AI framework** for precision agriculture, integrating **feature-based ML** and **image-based DL** for robust crop management. With high accuracy across all modules and a farmer-friendly interface, it represents a significant advancement toward **Agriculture 6.0**.

IX. Future Work

- Extend dataset to **field-level images** beyond PlantVillage.
- Integrate **satellite and IoT sensor data** for spatio-temporal predictions.
- Develop **mobile application** with offline support and regional language interfaces.
- Explore **transformer-based vision models (ViT, Swin)** for disease detection.
- Incorporate **time-series models (LSTMs, GRUs)** for yield forecasting.