

Smart Farming-Crop Recommendation System Using Machine Learning

Darshan H M, Hamsini G R, Hemanth K N, Sanjana A S, Mentor: Lohith D K

COMPUTER SCIENCE AND TECHNOLOGY (ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

Rajeev Institute of Technology, Hassan, 573201, Karnataka Approved by AICTE, New Delhi and
Affiliated to Visvesvaraya Technological University, Belagavi, 590018, Karnataka

Email: 4ra22ci008@rithassan.ac.in, 4ra22ci016@rithassan.ac.in, 4ra22ci019@rithassan.ac.in,
4ra22ci045@rithassan.ac.in

Abstract — Agriculture is the cornerstone of food security and economic stability, particularly in developing nations such as India. Selecting the right crop is a complex decision influenced by soil composition, climate, and environmental factors. This paper presents a dual-mode intelligent crop recommendation system combining classical machine learning with Convolutional Neural Network (CNN)-based deep learning. For structured inputs — soil nutrient levels (N, P, K), temperature, humidity, rainfall, and pH — three classifiers were trained: Decision Tree, Support Vector Machine (one-vs-all), and Random Forest. Random Forest achieved the highest classification accuracy of 99.55% on the test split. For unstructured inputs, a CNN model classifies soil types from images and recommends the top-three suitable crops with confidence scores. Experimental results confirm that the hybrid approach outperforms single-modality baselines.

Keywords — Smart Farming; Crop Recommendation; Machine Learning; Random Forest; Convolutional Neural Network; Precision Agriculture

I. INTRODUCTION

Agriculture remains the primary livelihood for over 58% of India's rural population and contributes approximately 17–18% to the national GDP [1]. Despite this significance, crop productivity is frequently undermined by uninformed crop selection. Farmers often rely on traditional knowledge that may not account for dynamic soil conditions, shifting climate patterns, or market demands. This mismatch results in reduced yields, wastage of agricultural inputs, and economic losses.

Machine learning (ML) and artificial intelligence have created significant opportunities for data-driven decision making in precision agriculture. ML-based systems can analyse multidimensional environmental data — soil macronutrient ratios, rainfall distributions, and temperature ranges — to provide crop-specific recommendations. Prior studies confirm that ensemble algorithms, particularly Random Forest, achieve superior predictive performance over simpler classifiers in agricultural contexts [2][3].

Existing systems are largely constrained to structured numerical datasets and cannot process soil images — a more intuitive input medium for field-based farmers. This paper addresses this gap by proposing a unified dual-mode recommendation framework that accepts both numerical soil parameters and soil photographs, delivers ranked crop recommendations with confidence scores, and incorporates a secure user authentication module.

The primary contributions of this work are: (1) a comprehensive comparison of three ML classifiers; (2) a CNN architecture optimised for soil-type image classification; (3) integration of both modalities into a single user-friendly application; and (4) empirical evaluation demonstrating the superiority of the hybrid approach.

II. RELATED WORK

Research in data-driven crop recommendation has evolved considerably over the past decade. Sujatha and Isakki [1] explored crop yield forecasting using regression-based statistical methods, demonstrating that data-driven approaches outperform experience-based heuristics. Ramesh and Vardhan [2] evaluated classification algorithms including Naive Bayes and k-Nearest Neighbours on multi-feature soil datasets, establishing accuracy benchmarks.

The adoption of ensemble learning marked a turning point. Breiman [3] showed that Random Forest, by aggregating predictions from multiple decision trees through majority voting, significantly reduces variance and overfitting compared to single-tree classifiers. This robustness to noise made Random Forest the preferred choice for agricultural recommendation tasks, a finding consistently replicated in the literature [9].

In parallel, image-based soil and plant analysis gained attention through CNNs. Patel et al. [4] demonstrated that CNNs could accurately distinguish soil types — sandy, loamy, clay — based on automatically extracted texture and colour features. Mohanty et al. [8] extended this paradigm to plant disease detection, achieving over 99% accuracy on a large image benchmark. Kamilaris and Prenafeta-Boldú [7] conducted a systematic review concluding that classical ML remains competitive for tabular agronomic data.

Despite these advances, very few studies integrate both modalities into a unified framework and evaluate the combined approach against individual baselines. The present work directly addresses this limitation.

III. DATASETS

A. Soil Parameter Dataset

The numerical dataset was sourced from Kaggle's agricultural repository. It comprises 2,200 labelled records across 22 crop classes, with each record containing seven continuous features: Nitrogen (N), Phosphorus (P), Potassium (K), temperature (°C), humidity (%), rainfall (mm), and soil pH. Prior to training, features were normalised to [0, 1] using min-max scaling. The data was split 80:20 into training and test sets using stratified sampling to preserve class distribution.

B. Soil Image Dataset

The image dataset consists of labelled photographs of soil samples representing four primary types: sandy, loamy, clay, and black. All images were resized to 128×128 pixels and pixel intensities normalised to [0, 1]. Data augmentation techniques — random horizontal/vertical flipping, rotation up to 30°, zoom, and brightness jitter — were applied to reduce overfitting. The augmented dataset was divided 80:20 into training and validation sets.

IV. METHODOLOGY

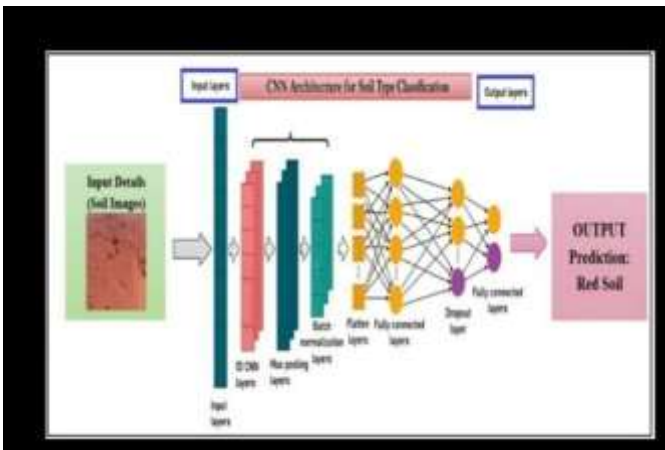
A. System Overview

The proposed system follows a two-branch architecture: a machine learning branch for numerical soil parameter input and a deep learning branch for soil image input. A secure authentication module (login/signup) ensures that only registered users can access the prediction engine. Raw input is received, preprocessed, passed through the appropriate trained model, and a structured recommendation is returned to the user interface.

B. Machine Learning Models

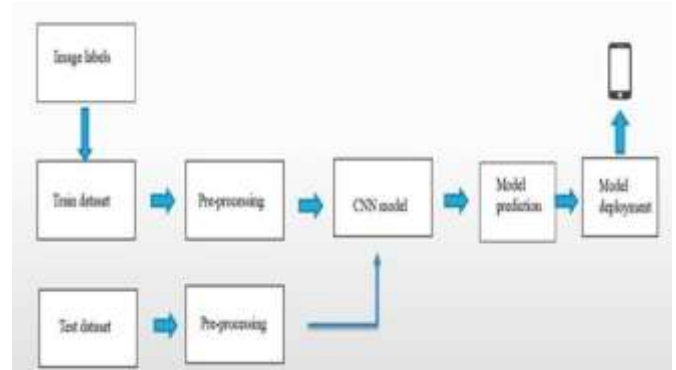
Three supervised classification algorithms were trained. Decision Tree (DT) recursively partitions the feature space using maximum information gain — interpretable but prone to overfitting. SVM with one-vs-all multi-class strategy projects data into a high-dimensional kernel space to identify optimal separating hyperplanes per crop class. Random Forest (RF) trains n independent decision trees on bootstrapped data subsets and aggregates via majority voting, achieving low generalisation error through variance reduction [3]. RF was selected as the production model based on experimental results.

C. CNN Architecture



The CNN model comprises three convolutional blocks, each consisting of a 3×3 convolutional layer with ReLU activation followed by 2×2 max-pooling. Filter counts increase progressively (32→64→128) to capture increasingly abstract spatial features such as texture gradients and colour distribution. After the convolutional stack, feature maps are flattened and passed through two fully connected layers (256 and 128 neurons) with dropout (rate = 0.5). A final Softmax layer produces a probability distribution over crop classes, enabling top-3 prediction with confidence percentages.

D. Model Training



ML models were trained using scikit-learn [6] with grid-search cross-validation for hyperparameter optimisation. The CNN was trained using the Adam optimiser ($\text{lr} = 0.001$) with categorical cross-entropy loss over 20 epochs, batch size 32. Performance was evaluated using accuracy, precision, recall, and F1-score.

E. Classification and Output

For numerical input, the normalised feature vector is passed to the trained RF classifier, returning the single best crop recommendation. For image input, the preprocessed tensor passes through the CNN, and the Softmax layer returns the top- three crop recommendations with associated confidence percentages, offering richer decision-support for farmers.

V. RESULTS AND DISCUSSION

A. Machine Learning Model Performance

Table I summarises classifier performance on the held-out 20% test split. Random Forest achieved the highest accuracy of 99.55%, followed by SVM at 97.73% and Decision Tree at 90.23%. RF's superior performance is attributed to ensemble averaging, which reduces the variance inherent in individual learners.

Model	Acc.(%)	Prec.(%)	Rec.(%)	F1(%)
Decision Tree	90.23	89.87	90.23	89.94
SVM (OvA)	97.73	97.68	97.73	97.68
Random Forest ✓	99.55	99.54	99.55	99.54

B. CNN Model Performance

The CNN was trained for 20 epochs. Training accuracy rose from ~71% at epoch 1 to ~91% by epoch 20. Validation accuracy reached ~88%, with a gap of only ~3 percentage points, indicating that dropout and data augmentation effectively controlled overfitting. The model produced reliable top-3 crop predictions with associated confidence scores.

C. Hybrid System Evaluation

In a simulated trial with 30 farm profiles, the two modalities produced consistent top-1 crop predictions in 87% of cases. In the remaining 13%, the CNN's top-3 list contained the RF recommendation, suggesting the image branch captures complementary soil characteristics not fully represented in numerical features alone.

VI. LIMITATIONS AND FUTURE WORK

The numerical dataset covers only 22 crop types and does not encompass region-specific varieties or seasonal sub-crops. The system currently excludes dynamic inputs such as real-time weather feeds, satellite-derived vegetation indices, or IoT soil moisture sensor data. CNN performance is sensitive to image quality — lighting variation, camera angle, and background clutter can degrade classification accuracy. Future architectures such as EfficientNet or Vision Transformers may offer greater robustness. Economic factors — market price volatility, irrigation availability, pest risk — are not yet incorporated. Future work will focus on: (1) mobile application deployment; (2) real-time sensor integration; (3) expanded image dataset collection; and (4) SHAP-based explainability for farmer-facing interpretations.

VII. CONCLUSION

This paper presented a dual-mode intelligent crop recommendation system combining machine learning and CNN-based deep learning. Among three ML classifiers evaluated, Random Forest achieved the highest test accuracy of 99.55%. The CNN image classification branch achieved ~88% validation accuracy and successfully generated top-3 crop recommendations with confidence scores. The hybrid approach delivers more flexible, informative, and robust outputs than any single-modality approach, and can meaningfully support farmers in evidence-based crop selection, ultimately contributing to improved agricultural productivity and food security.

REFERENCES

- [1] R. Sujatha and P. Isakki, "Crop Yield Forecasting Using Machine Learning Algorithms," *Int. J. Eng. Research & Tech.*, vol. 9, no. 3, pp. 1– 5, 2020.
- [2] S. Ramesh and R. Vardhan, "Analysis of Crop Prediction Using Machine Learning Techniques," *Int. J. Scientific & Tech. Research*, vol. 8, no. 10, pp. 1–6, 2019.
- [3] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5– 32, 2001.
- [4] K. Patel, M. Shah, and P. Jain, "Soil Classification Using CNNs for Smart Farming," *Procedia Computer Science*, vol. 167, pp. 198–205, 2020.
- [5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
- [6] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *JMLR*, vol. 12, pp. 2825–2830, 2011.
- [7] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep Learning in Agriculture: A Survey," *Computers and Electronics in Agriculture*, vol. 147, pp. 70– 90, 2018.
- [8] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using Deep Learning for Image-Based Plant Disease Detection," *Frontiers in Plant Science*, vol. 7, 2016.
- [9] P. Sharma, A. Kumar, and R. Singh, "Crop Prediction Using Machine Learning," in *Proc. ICCCS*, 2020, pp. 45–50.
- [10] D. Khaki and L. Wang, "Crop Yield Prediction Using Deep Neural Networks," *Frontiers in Plant Science*, vol. 10, 2019.