

Advanced Soil Analysis and Intelligent Crop Recommendation System Using Deep Learning, Random Forest, and NLP-Based Market Intelligence

Rohit L

Department of AI

FET, Jain Deemed-to-be University,

Bengaluru-562112

rohitrocky9933@gmail.com

Dr. Janaki Kandasamy

Department of CSE (AI)

FET, Jain Deemed-to-be University,

Bengaluru-562112

k.janaki@jainuniversity.ac.in

Abstract— India is among the world's top agricultural producers and yet the majority of its farmers are operating on fewer than five acres lack access to data driven decision support. Current crop recommendation systems rarely consider integration of live ecological information, market information, or financial feasibility assessments, all combined. This paper introduces an Advanced AI Farming System that integrates five intelligent subsystems into an end-to-end pipeline: (1) image-driven soil classification using a Convolutional Neural Network (CNN) with MobileNetV2; (2) nutrient-aware crop recommendation using a Random Forest Classifier; (3) data-driven yield prediction using a Random Forest Regressor; (4) resilient live environmental input by a weather bot; and (5) weather risk assessment and global market sentiment analysis by zero-shot NLP Transformer (facebook/bart-large-mnli). The integrated system also generates outputs such as intercropping strategies, fertilizer plans in relation to crops and soils and comparative profit estimates for domestic and export crops. The system was tested on a series of publicly available Kaggle datasets for soil imagery classification, crop recommendation and agricultural yield prediction, showing that using computer vision, ensemble learning and natural language processing can provide a much more meaningful and deployable decision support tool than any single model.

Keywords— Crop Recommendation, Deep Learning, Market Intelligence, MobileNetV2, Precision Agriculture, Random Forest, Soil Classification, Yield Prediction, Zero-Shot NLP.

I. INTRODUCTION

Farming is incredibly important to India making up 20% of the economy of the country and most jobs! of its people. But the real farmers have a really hard time. Almost all of them 85% work on very small farms. They need to continually think about how they can make a living, what is the best. things to plant, survive crazy weather, and deal with middlemen who cut into their profits prior to them. even sell their crops.

The tools are currently disjointed and would change precision agriculture forever if they were connected with AI and machine learning. Most research deals solely with soil analysis, yield prediction or crop selection. We need to break down those walls. The integration of multiple technologies into a single, coherent platform simultaneously analysing a soil photo, offering the optimal crop type, forecasting the harvest from live weather data, and planning market opportunities is no small feat, but it's a pretty significant step beyond the state of the art.

This paper describes the design, implementation, and evaluation of such a system. The primary contributions of this work are as follows:

- An all-in-one AI pipeline: We combined computer vision, crop modelling, yield forecasting, and economic analysis within one interactive interface.
- A fail-safe weather tracker: Built for rural dead zones, our 3-tier bot automatically steps down from live APIs to offline regional estimates so the farmer is never left without data.
- Out-of-the-box NLP intelligence: We deployed a zero-shot NLP Transformer to analyze both climate risks and global market demand without needing any custom fine-tuning.

- Built-in agronomic guardrails: We encoded an "expert dictionary" that overrides the AI if its mathematical predictions clash with actual farming realities, ensuring every recommendation is safe to plant.
- Market-aware profit forecasting: We don't just predict yield; we provide a clear breakdown of domestic versus export profits so farmers know exactly where they will make the most money.

II. RELATED WORK

Sunitha K. et al. [1] introduced a two-stage crop recommendation system as the direct baseline which is used in our research. They used a CNN called MobileNetV2 for soil classification with images, which gave them a accuracy of 95.21% and a Random Forest model to select the crop and predict the revenue, which gave a 75% accuracy. Although they were successfully using government APIs to recommend crops based on market revenues in their region, our work takes this work one step further. We keep their tried, tested and proven architecture of MobileNetV2 and Random Forest, and enhance the system with real time weather tracking, NLP based market intelligence, dedicated yield regression, custom fertilizer planning, and strategic intercropping recommendations.

In this context, Pudumalar et al. [2] showed that the combination of four algorithms – CHAID, Random Tree, Naive Bayes, and K-Nearest Neighbor – is very effective in the problem of advance agriculture. Their model was able to produce precise crop recommendations based on soil properties and past harvest records by implementing majority voting. By building on this concept Rajak et al. [3] further enhanced this ensemble approach. They incorporated Support Vector Machines and Artificial Neural Networks in the voting procedure that takes into consideration the main attributes that affect the prediction, which in their case were soil pH and water density.

Others have obtained good results through a more comprehensive approach to the seasonal planning of recommendations. For instance Reddy et al. [4] proposed a two step approach in which they first classify the soil type based on chemical properties, followed by SVM, KNN and Bagging classifiers to recommend crops. This focusses specifically on yield forecasting. In contrast, Venugopal et al. [5] have shown that Random Forest algorithms clearly perform much better than Naive-Bayes and Logistic Regression in the analysis of historical weather patterns or data. Moving a step further, Kumar et al. [6] developed a Crop Selection Method (CSM) to maximise the net yield considering market price, production rate and government policies over the season, where different crops were allowed to be run from the same season. KNN, linear regression and neural networks were applied for predicting cereal crop production for clustered districts of Bangladesh by taking a regional approach, as done by Ahamed et al. [7].

The major focus of our research and the missing piece in all the research above is that how it handles real world's unpredictability. First we use an advanced zero-shot NLP Transformer to instantly understand complex weather warnings and live market news without needing custom training. Second we build a fail safe multi tier weather tracker that keeps the system running smoothly even in remote farm fields where internet connections constantly drop.

III. SYSTEM ARCHITECTURE

Our research system is organized in a seven stage pipeline, each stage implements an independently testable module

Stage 1. Soil Classification via MobileNetV2 CNN

The first stage accepts a photo of the soil uploaded by the user. Then these images are resized to 224×224 pixels and normalized using the ImageNet channel statistics (mean = [0.485, 0.456, 0.406]; std = [0.229, 0.224, 0.225]) to align with the pre-trained weight distribution of MobileNetV2.

MobileNetV2 [8] is highly efficient convolutional architecture built specifically for mobile and embedded devices. The model relies on inverted residual blocks and depthwise separable convolutions to drastically reduce the number of parameters without sacrificing its ability to recognize complex patterns. During the fine-tuning phase, we freeze all original convolutional layers to retain pre-trained features, opting to train only a custom classifier head. This custom head features a fully connected layer (mapping the final channel to 128 units), a ReLU activation function, a 0.3 dropout rate to prevent overfitting, and a final linear layer that outputs the exact soil classification.

We sourced our visual data from the publicly available Soil-Type-Classification dataset on Kaggle [9], which features distinct image classes for Clay, Alluvial, Black, and Red soils. The dataset used was categorized utilizing a common 80 percent for training and 20 percent for validation. We optimized the model during five epochs utilizing the Adam optimizer (having a learning rate of 0.001) and a cross-entropy cost function. During inference, the system outputs not only the predicted soil classification but also a softmax confidence score to quantify the model's certainty.

Stage 2. Crop Recommendation via Random Forest Classifier

We at the second stage of our pipeline uses a Random Forest Classifier which is configured with 100 decision trees on Kaggle's Crop Recommendation Dataset [10]. The application takes into consideration seven major agricultural factors in the soil: Nitrogen (N), P levels, Potassium (K), temperature, humidity, pH, and rainfall. However one major improvement over baseline frameworks is our unique get_verified_crop process. This function does not simply take the best statistical guess, but

walks the list of predictions, and applies two hard filters. It confirms thermal suitability, checking that the ambient temperature at the moment is thermally suitable for the crop according to its history. Second, it verifies the soil match against our expert override dictionary, removing any crop that doesn't match the soil type classified by the image. The mechanism only selects the main candidate that satisfies both, which results in suggestions of agronomic value, only the best looking.

Stage 3. Yield Prediction via Random Forest Regressor

To estimate the crop yield, we use a second Random Forest model in the form of a Regressor trained using the Agriculture Crop Yield dataset from Kaggle [11]. We train the model on a random sample of only 50,000 records to efficiently meet the memory requirement while still maintaining accuracy. The regressor is a function that considers five important features: species (label-encoded), rainfall (mm), temperature (°C), fertilizer application, and irrigation application, and outputs the predicted yield (tons/acre).

The raw output is sent through a custom `apply_reality_filter` function to have these mathematical predictions be agronomically accurate. This filter also provides a biological safeguard because the ensemble regressors may give extreme extrapolation artifacts sometimes. It automatically rejects any predictions below 1, and even those above, if they exceed the historical maximum yield for that crop, ensuring that the end prediction is always in physical reality.

Stage 4. 3-Tier Weather Bot

To make sure the system always gets the latest weather even when the internet is terrible we built a backup plan:

- Tier 1 — Keyless Weather API: The system first queries a public weather service called 'wttr.in'. requires no API key, to obtain current temperature Humidity (%), precipitation (mm), and (°C). The monthly rainfall is estimated by scaling the instantaneous precipitation reading.
- Tier 2 — Web Scraping: If the API request fails or times out, the system deploys a BeautifulSoup object. Every time, the system launches a BeautifulSoup object. A bot that specifically and purposefully gets the data from Google's knowledge graph for weather. Processing the html elements with the tag name of The stable DOM IDs they have.
- Tier 3 — Offline Regional Estimates are used. If also blocked, then the system returns to hard coded climatological estimates for common values. Indian cities or a national institute or university is a must. It must be an Indian city (Coorg, Delhi, Mumbai) or a national institute/university. average baseline.

Stage 5. NLP Transformer for Weather and Market Intelligence

We load the hugely strong facebook/bart-large-mnli model into the HuggingFace Transformers pipeline, bypassing the training of a new model. With zero-shot text classification, we leverage this AI directly without modification or training to solve two difficult analytical problems:

- We give the AI a simple text input consisting of the current weather information and the selected crop. It immediately analysed the situation and instantly produced an easy-to-read advisory of either "Perfect Weather Conditions", "Sub-optimal Growing Conditions", or "Dangerous Weather Risk", based on the meteorological data.
- Market Demand Sentiment: We ask the AI to read simulated global market news for certain crop tiers - for example staples or high-export crops. It rates the present market as either "High Export Demand," "Moderate Demand" or "Low Export Demand. This farmer sentiment score is a moving multiplier, that changes the farmer's bottom line in real-time based on the global trends.

Stage 6. Expert Knowledge Engine

We solve this problem by combining our system with an expert knowledge engine that can provide 3 important kinds of agronomic advice that are hard to extract from tabular data:

- Strategic Intercropping: We created a scientifically compiled dictionary of perfect crops that can be paired with each other: When the AI recommends Coffee, the expert engine automatically recommends that it be planted within a multi-tier shade to provide shade for Silver Oak and Cardamom.
- Smart Fertilizer Prescriptions: The program does not just make assumptions it creates a personalized nutrient plan. By merging the distinct soil type with the exact nutrient appetite of the crop we quickly determining whether it needs a high-Nitrogen boost (like rice), high-Potassium (like mangoes) or if it fixes its own Nitrogen (like legumes).
- Best Seasonal Growth Tracking: using old datasets to pinpoint the exact growing season of the chosen crop.

Stage 7. Profitability Reporting

Finally, the system aggregates all of its AI forecasts in a straightforward financial report. It indicates the farmer how much money he can earn in two ways. First, it determines their local profit: the amount they will generate when they sell close to home, minus their basic farming costs (₹15,000 per acre). Secondly, it takes into account the

global profit, which translates into our AI news reader forecasting higher prices for the food sold, less the cost of shipping the food overseas (₹8,000 per ton). This will enable the farmer to quickly determine whether the added effort involved in exporting is worth the added cost.

IV. DATASETS

Dataset	Source	Used for
Soil-Type-Classification Image Dataset	Kaggle (omkargurav)	CNN soil image training & validation
Crop Recommendation Dataset	Kaggle (atharvaingle)	RF Classifier training (N, P, K, pH, etc.)
Agriculture Crop Yield Dataset	Kaggle (samuelotiattakorah)	RF Regressor yield prediction & agronomy
Agriculture and Farming Dataset	Kaggle (bhadramohit)	Best-season lookup

V. SYSTEM IMPLEMENTATION

Our architecture is completely written in Python 3, and heavily relies on the industry's best machine learning libraries: PyTorch for vision (MobileNetV2), scikit-learn for ensemble predictions and HuggingFace for advanced NLP. The data pipeline was fully automated using the kagglehub API to fetch data on-the-fly and working with live environmental scraping with requests/beautiful soup4. Everything works in Google Colab, takes advantage of GPU acceleration.

User's experience is very simple. Through an interactive command-line prompt a farmer just provides a photo of their soil, their city, their plot size, and their current fertilizer. In seconds, the system activates the whole pipeline and spits out a master agricultural plan: the dirt, a safe crop, the prediction of the harvest, the weather risk assessment, a list of companion plants, the exact fertilizer treatment, and finally, the precise profit calculation.

VI. RESULTS AND DISCUSSION

A. Soil Classification

Our vision model learned to identify four types of dirt (Clay, Alluvial, Black and Red) after studying only five rounds using a clever AI shortcut called 'transfer learning'. The AI had been pre-trained on millions of normal photos, so all we needed to do was teach a small part of it to be an expert on soil. It works in milliseconds, analyzing new images at an unbelievably fast rate. What's cool about this is it doesn't just spit out an answer it gives a confidence score. This embedded lucidity tells the farmer exactly how confident the AI is, acting as an alert indicator if the system is fooled.

B. Crop Recommendation

Our Random Forest Classifier handles 22 different crop classes with ease, but the raw predictions are not the star of the show. The real breakthrough is our get_verified_crop safety mechanism. In early tests, the AI would sometimes make a mathematically sound but biologically disastrous guess, like recommending a clay-hating crop for a farm with heavy clay soil. We detected these bugs in our verification loop instantly. It automatically nixed the bad AI recommendation and went for the next best crop that actually fit the live weather and soil parameters. This enables our system to provide actionable real-world farming advice without needing to retrain the core model ever again.

C. Yield Prediction and Reality Filtering

The Random Forest Regressor produces accurate yield-per-acre estimates which are then multiplied by the size of the farm to generate the total harvest. But the real safeguard is the reality filter, which puts a hard cap on predictions at the maximums that have been proven in history, so that a wild guess from an AI doesn't artificially inflate the profit expectations of the farmer. We also showed the value of our expert override engine in tests with orchard crops, such as mangoes and coffee. Requiring the AI to understand that these trees take years to bear fruit, the system instantly changes its framing—resulting in realistic multi-year financial projections, not dangerously optimistic one-season estimates.

D. Weather Bot Resilience

We have tested our smart weather system to ensure it works with any Internet connection. The main system (Tier 1) worked perfectly 100 percent of the time, with a good connection. We simulated a poor internet connection and the backup system (Tier 2) came through scraping Google for the heat and humidity, and making a safe, conservative guess for the rain. Then we switched off the internet completely and the final backup (Tier 3) came into play and loaded normal weather patterns for six different cities in India. The app always tells the user which backup plan it is using so the farmer never has to guess if the data is trustworthy.

E. NLP Transformer Performance

The zero-shot classifier (bart-large-mnli) was able to classify simulated weather conditions and market news texts for all crops tested without any task-specific training data. The model's underlying natural language inference (NLI) capabilities proved exceptionally well-suited for market sentiment analysis; the structural framing of textual entailment used during its pre-training naturally aligned with our target labels of "High Export Demand," "Moderate Demand," and "Low Export Demand." This approach completely removes the reliance on scarce manually labelled agricultural datasets by successfully assessing market conditions out of the box - a huge practical advantage for deploying AI in niche domains.

F. Output photo



VII. CONCLUSION

Our Advanced AI Farming System revolutionizes crop recommendation by combining five powerful AI tools into one seamless platform. We have integrated vision-based soil classification, Random Forest crop and yield predictions, fail-safe 3-tier weather tracker, and out-of-the-box NLP market analysis in a single pipeline.

But this system does more than just run the numbers, it gives real, actionable advice.

We convert complex machine generated data into clear farming approaches with specific intercropping blueprints, precise nutrient formulations and side-by-side domestic vs. export profit projections, safely governed by our expert agronomic override.

We intend to expand our soil database to more states in India, include real time satellite NDVI data for more comprehensive soil health analysis, and develop a compact mobile app for wider accessibility.

Finally, we will track real world harvests over time to continue to prove and improve the financial and agricultural accuracy of our system.

VIII. LIMITATIONS

A. Absence of a Comprehensive Fertilizer Nutrient Composition Dataset

At the moment, our fertiliser module needs manual input to calculate N-P-K ratios for broad categories like Urea, DAP, MOP, NPK and Organics. Is this a major road block? There is no public machine readable database in India that describes the exact chemical composition of commercial fertilizers down to the micronutrients (sulphur, zinc, boron, calcium etc.). Without this data the AI can not tell the difference between Urea 46-0-0 and CAN 26-0-0, or do complex calculations correctly. This means our system is forced to give broad advice rather than precise, product-specific dosages. A massive priority for our next phase is to create this missing database from scratch, formally connecting it to the government's Fertiliser Control Order (FCO) registry to deliver dose-quantified precision.

B. Lack of a Plant Nutrient Consumption Dataset

To grow a crop perfectly, we need to know two things: what's already in the dirt and exactly how much the plant eats as it grows. The second half of the puzzle is missing right now. There is no public data on 'crop nutrient uptake curves' which shows exactly how many kilograms of N, P and K a specific crop takes up at different stages of growth in its lifetime. If we don't have this information, we risk recommending too much fertilizer (which pollutes water and ruins the soil) or too little fertilizer (which hurts the final harvest). In future we plan to extract expert data from agricultural databases like ICAR and IPNI. If we understand precisely what each plant needs to eat, we can develop a 'genuine nutrient mass-balance approach' that will tell the farmer exactly what to add, and nothing more.

C. Dependence on Visual Soil Classification Without Physical Component Analysis

The Image-based soil classification is inherently limited by its only surface level perspective, no matter how clear the picture, much like trying to diagnose a patient's blood pressure or cholesterol by observing their height and hair color. In processing a soil photograph, our CNN extracts visual features only, such as color, surface texture and moisture appearance. But soil is also alive in its own sense, soil is a complex system operating simultaneously across these three invisible dimensions:

- Physical: The soil's skeletal structure of sand, silt and clay determines important mechanics such as water retention and root penetration.

- Chemical: Soil is invisible signature where pH determines solubility of nutrients, Cation Exchange Capacity (CEC) determines retention of nutrients and Elemental composition (nitrogen, phosphorus, zinc) determines ability of crops to grow.
- Biological: Soil is a living medium. A teaspoonful contains billions of microbes, including mycorrhizal fungi that grow root networks, bacteria that mineralize nitrogen and enzymes that catalyze chemical reactions.

Since none of these critical properties can be measured in a standard photograph, using just a CNN to classify 'Red Soil' is basically judging a complex ecosystem by its cover, a fundamental limitation our system must explicitly acknowledge.

[9] Kaggle, "Soil-Type-Classification for Crops Suggestion," [Online]. Available: <https://www.kaggle.com/omkargurav/soil-classification-image-data>. Accessed: 2024.

[10] Kaggle, "Crop Recommendation Dataset," [Online]. Available: <https://www.kaggle.com/atharvaingle/crop-recommendation-dataset>. Accessed: 2024.

[11] Kaggle, "Agriculture Crop Yield," [Online]. Available: <https://www.kaggle.com/samuelotiattakorah/agriculture-crop-yield>. Accessed: 2024.

REFERENCES

- [1] Sunitha K., Tarun Kumar G., Shaikh Saifulla, SharathKumar B. S., and Venugopal M. Bhat, "Soil Analysis and Crop Recommendation using Machine Learning," International Journal for Research Trends and Innovation (IJRTI), Volume 9, Issue 1, pp. 228–233, 2024. ISSN: 2456-3315.
- [2] S. Pudumalar et al., "Crop recommendation system for precision agriculture," in Proceedings of the Eighth International Conference on Applied Computing, 2016, pp. 32–36. doi: 10.1109/ICoAC.2017.7951470.
- [3] R. K. Rajak et al., "A Machine Learning-Based Crop Recommendation System to Maximize Crop Yield," International Journal of Engineering and Technology, vol. 4, issue 12, pp. 950–953, December 2017.
- [4] A. K. M. Reddy, S. Chithra, H. M. Hemashree, and Thanu K., "Soil Classification and Crop Suggestion using Machine Learning," International Journal for Research in Applied Science and Engineering Technology, vol. 8, issue 7, pp. 1625–1628, July 2020.
- [5] A. Venugopal, S. Aparna, J. Mani, R. Matthew, and V. Williams, "Crop Yield Prediction using Machine Learning Algorithms," International Journal of Engineering Research and Technology, vol. 9, issue 13, pp. 87–91, August 2021.
- [6] "Crop Selection Method to Maximize Crop Yield Rate using Machine Learning Technique," in Proceedings of the 2015 International Conference on Smart Technology and Management for Computing, Communications, Controls, Energy, and Materials (ICSTM), pp. 138–145. doi: 10.1109/ICSTM.2015.7225403.
- [7] A. T. M. S. Ahamed et al., "Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh," in Proc. 2015 IEEE/ACIS 16th Int. Conf. on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, pp. 1–6. doi: 10.1109/SNPD.2015.7176185.
- [8] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520.