**1. Overview**

Crop failure remains a major concern for farmers and agricultural economies, especially in developing countries like India. Traditional methods of monitoring are often reactive, inaccurate, or delayed. With the rise of AI and access to multivariate datasets (like weather, soil, crop, and satellite data), there is a growing opportunity to build smarter systems that can predict failure before it happens. This paper proposes a beginner-level AI framework that combines various data inputs to improve the prediction of crop failure, supporting farmers and policymakers with early alerts and insights.

**2. Introduction**

Agriculture is highly dependent on natural and environmental factors. Unpredictable rainfall, poor soil quality, excessive pesticide use, and emerging plant diseases can all contribute to crop failure. Although some AI solutions have been developed in this area, most are limited to one or two parameters — such as only using rainfall or NDVI values. This study aims to build a multivariate crop failure prediction system using weather, soil, satellite imagery, and plant health data. The model, trained with diverse data sources, will offer farmers and authorities a better way to anticipate failure risks and take early preventive measures

**3. Problem Statement**

Most existing crop prediction models:

* Focus on a **single data type** (e.g., only weather or only NDVI)
* Lack real-time adaptability
* Do not incorporate environmental degradation like **deforestation, air pollution**, or **pest outbreaks**
* Are often **too complex or inaccessible** to grassroots-level users

As a result, these systems fail to accurately reflect the real conditions that lead to crop failure. There is a need for a **comprehensive and accessible AI model** that includes multiple real-world factors and delivers early predictions with higher accuracy.

**4. Proposed Solution (Our Approach)**

We propose building an AI-based predictive model that integrates multiple types of data to detect and predict the likelihood of crop failure in a region. The system will:

* Collect open-access data from reliable platforms (IMD, ISRO, NASA, etc.)
* Train ML algorithms (starting with Random Forest and Logistic Regression)
* Analyze patterns leading to historical crop failures
* Evaluate the model using accuracy, F1 score, and confusion matrix
* Provide a visual dashboard or alert system for easy interpretation

This research will serve as a prototype that can be expanded and optimized in later stages using deep learning or real-time feedback systems.

**5. Factors We Will Consider**

**📊 Environmental Factors**

* Rainfall
* Temperature
* Humidity
* Air quality
* Pollution (soil/water)

**🌱 Soil & Crop Factors**

* Soil pH
* NPK levels (Nitrogen, Phosphorus, Potassium)
* Crop type
* Crop season
* Overuse of fertilizer/pesticides (from manual/local reports)

**🛰️ Satellite & Remote Sensing**

* NDVI (vegetation health)
* Drought index
* Surface temperature
* Evapotranspiration
* Deforestation indicators

**🐛 Pest & Disease Indicators**

* Historical outbreaks
* Crop disease maps
* Pest infestation patterns (if available)

# Data type collection

**🌦️ 1. Weather Data**

**✅ Source:**

[**India Meteorological Department (IMD)**](https://mausam.imd.gov.in/)

* Rainfall, temperature, humidity
* Past weather reports & crop season data (may require request for historical data)

**✅ Alternative:**

* [NASA POWER](https://power.larc.nasa.gov/) (Global weather & solar data)
* [OpenWeather API](https://openweathermap.org/) – for current weather (limited free tier)

**🧪 2. Soil Data**

**✅ Source:**

[**Soil Health Card India**](https://soilhealth.dac.gov.in/)

* pH, Nitrogen, Phosphorus, Potassium
* District/block-wise reports

**✅ Alternative:**

* [ISRO BHUVAN Soil Map](https://bhuvan.nrsc.gov.in/home/index.php)
* FAO SoilGrids: [soilgrids.org](https://soilgrids.org/)

**🌱 3. Crop Type, Season & Yield**

**✅ Source:**

[**Directorate of Economics and Statistics – Ministry of Agriculture**](https://eands.dacnet.nic.in/)

* Crop-wise and state-wise productivity data
* Historical yield reports

**🛰️ 4. Satellite Data (NDVI, Deforestation, Surface Temp)**

**✅ Source:**

[**Google Earth Engine (GEE)**](https://earthengine.google.com/)

* NDVI (vegetation health index)
* Surface temperature
* Land cover change / deforestation
* Evapotranspiration

💡 GEE also lets you write simple JavaScript code to extract satellite data as CSV.

**✅ Alternative:**

* [NASA FIRMS](https://firms.modaps.eosdis.nasa.gov/) (fire and heat maps)
* [ISRO Bhuvan](https://bhuvan.nrsc.gov.in/) – NDVI, drought, land use data

**🐛 5. Disease & Pest Outbreaks**

**✅ Source:**

[**CROPDATA from CIBRC**](https://ppqs.gov.in/divisions/cib-rc)

**or**

[**ICAR**](https://icar.org.in/)

* Check bulletins and pest/disease alerts
* ICAR Research Centers (state-wise) may provide data on request

**🌫️ 6. Pollution & Air Quality**

**✅ Source:**

[**CPCB (Central Pollution Control Board)**](https://cpcb.nic.in/)

* Air quality, water contamination
* May require zone-wise manual data gathering

**🔍 7. Datasets for Training ML Models (Already Cleaned)**

* [Kaggle – Crop Yield or Failure Datasets](https://www.kaggle.com/search?q=crop+failure)
* [UCI Machine Learning Repository](https://archive.ics.uci.edu/)
* [Open Data India](https://data.gov.in/)

**✅ Bonus: Combine multiple data sources**

**SO LET’S DIVIDE**

**🎯**

**Goal**

**: Build a system to predict**

**crop failure using AI**

**, with weather, soil, crop, pest, and satellite data.**

**🔧**

**Part A: Person 1 – Data Science & AI (Model Builder)**

**🟢 Beginner Skills:**

* ✅ Python basics
  + Variables, loops, functions, conditionals
  + Libraries: pandas, numpy, matplotlib
* ✅ Data cleaning & preprocessing
  + Handle missing values, outliers, normalization
* ✅ Data visualization
  + seaborn for correlation plots, heatmaps, graphs

**🔵 Intermediate Skills:**

* ✅ Machine learning (with Scikit-learn)
  + Logistic Regression, Decision Tree, Random Forest
  + Model training, testing, evaluation (accuracy, confusion matrix, F1 score)
* ✅ Feature engineering
  + Create custom features (e.g. rainfall deviation, NDVI index category)
* ✅ Model tuning
  + Grid search, cross-validation, overfitting/underfitting

**🔴 Advanced Skills (Later stage):**

* 🧠 Deep Learning (TensorFlow/PyTorch) — optional
* 📈 Time series forecasting (for yield prediction)
* 🧰 Model explainability (SHAP, LIME)
* 🧪 Deploy model via Streamlit or Flask (if app version is needed)

**🛰️**

**Part B: Person 2 – Data Gathering, Satellite, and Writing (Researcher + Backend)**

**🟢 Beginner Skills:**

* ✅ Python basics
  + File handling, data loading, CSV reading/writing
* ✅ Web scraping basics
  + Using BeautifulSoup or Selenium to gather open data
* ✅ Google Sheets or Excel
  + Organize large datasets, clean manually if needed

**🔵 Intermediate Skills:**

* ✅ Google Earth Engine (GEE)
  + NDVI, evapotranspiration, land cover
  + Export image stats to CSV
* ✅ QGIS or GEE Explorer
  + Visualize remote sensing data
* ✅ Satellite image understanding
  + Vegetation health, drought zones

**🔴 Advanced Skills (Optional):**

* 🛰️ Image classification (NDVI thresholding, CNN use)
* 🌍 GIS programming (geospatial Python like geopandas, rasterio)
* 🗃️ Build a database (SQLite or simple cloud storage for big data)

**📜 Both Will Need (Research & Writing):**

* 📚 How to read research papers (Google Scholar, Semantic Scholar)
* 📄 Structure of a research paper (Abstract, Problem, Methodology, Result)
* ✍️ Writing tools: MS Word / Google Docs / Overleaf (LaTeX optional)
* 📖 Referencing & citation with **Zotero** or **Mendeley**

**✅ TL;DR – Your Combined Learning Plan**

|  |  |  |
| --- | --- | --- |
| Area | Person 1 (AI Dev) | Person 2 (Data + Writing) |
| Python & Pandas | ✅ | ✅ |
| ML/Scikit-learn | ✅ (focus area) | ❌ |
| Satellite/GEE | ❌ | ✅ (focus area) |
| Data Collection | ✅ | ✅ |
| Paper Writing | 🟡 | ✅ |
| Web Scraping | ❌ | ✅ |
| Model Deployment | 🟡 | ❌ |

Great! Let’s make sure the **data collection workload is split evenly** between both of you — based on both **importance** and **learning opportunity**. This way you both get hands-on experience with a mix of technical and research tasks.

**✅ Balanced Data Collection Task Split**

|  |  |  |
| --- | --- | --- |
| Data Type | Description | Assigned To |
| 1. Weather Data | Rainfall, temperature, humidity (from IMD, NASA POWER) | 👤 **Person 1** |
| 2. Soil Data | Soil pH, NPK levels (Soil Health Card, ISRO Bhuvan) | 👤 **Person 2** |
| 3. Crop Type & Season Data | Crop info, sowing/harvest season, cropping pattern | 👤 **Person 1** |
| 4. Pest/Disease Reports | Reports from ICAR, CIBRC, news sources, government alerts | 👤 **Person 2** |
| 5. NDVI & Satellite Imagery | NDVI, land cover, drought zone (via Google Earth Engine) | 👤 **Person 2** |
| 6. Deforestation Data | Land use change, satellite image comparison (ISRO, NASA) | 👤 **Person 1** |
| 7. Air & Water Pollution | AQI, chemical contamination in agri zones (CPCB, state data portals) | 👤 **Person 2** |
| 8. Crop Failure Labels | Create or extract labels from low-yield or severe-weather reports | 👤 **Person 1** |
| 9. Dataset Integration | Clean, merge, and align all data into a final dataset | 👥 **Both together** |

**🧠 Smart Swap Strategy:**

* Person 1 is more AI/model-focused → gets structured & label-heavy datasets
* Person 2 is more research/satellite-focused → gets imagery & field-specific data
* **Final dataset creation** should be a **collaborative step** since it connects everything.