Project Track Chosen: Crop Monitoring - AI Tools for Growth Analysis and Yield Prediction

Project Overview:

Agriculture remains a key pillar of the Indian economy, but farmers face increasing uncertainty due to fluctuating weather, market volatility, and lack of reliable data on crop health and expected yield. This project aims to build an intelligent system to predict crop yield accurately using machine learning models, while ensuring transparency and trust via Explainable AI (XAI) techniques.

A major innovation in this project is the first-time use of the CY-Bench dataset in an India-focused system, bringing together standardized, multi-modal sub-national crop data to develop highly robust prediction tools. Alongside, the system also integrates NDVI-based crop monitoring, weather and soil data making it a full solution from most recent problems faced in recent research in this yield prediction.

Solutions We Proposed are:

1. Crop Yield Prediction Model

- **Dataset**: Uses CY-Bench, which includes:
    - **NDVI**, **FPAR**, **weather data** (temperature, rainfall, radiation)
    - **Soil moisture** and **static soil features**
    - **Crop calendars** and **annual yield data**
- **Feature Engineering**: Raw satellite and ground data is pre-processed to extract informative features.
- **Model Training**: Multiple ML models will be trained and tested and best model choosing:
    - **Hyperparameter tuning** to optimize accuracy.
    - **Evaluation metrics** like R^2, MAE, and RMSE.
- **Explainable AI (XAI)**: Helps visualize what features influenced the yield prediction, building trust and transparency.

2. NDVI-Based Crop Monitoring

- **User Input**:
    - Users can select a region using **latitude/longitude** or a **Google Maps tool**.
    - Maximum monitored area is **5 km²**.
- **Data Sources**:
    - **Bhuvan NRSC** for India-specific NDVI.
    - **Sentinel or Landsat imagery** via Google Earth Engine.

- **Purpose**:
    - Real-time crop health tracking using NDVI trends and anomaly detection.

3. Weather and Soil Data Integration

- **API Used**: Open Meteo API
- **Data Features**:
    - Forecasted and historical temperature, rainfall, soil moisture and many more data available.
- **Use**:
    - Enhances yield prediction by factoring in environmental conditions.

4. CY-Bench Dataset Highlights:

- Multi-modal, sub-national dataset available in time-series CSVs.
- **Covers:**
    - NDVI, FPAR, temperature, rainfall, radiation
    - Soil moisture, static soil properties
    - Yield, production, and crop calendars
- **Spatial Data:**
    - Includes region shapefiles with centroids and boundaries.
- **Advantages:**
    - Clean, standardized benchmark for training robust and generalizable models.

Team Contributions:

Name	Contribution
Neeraj Jaiswal	Conducted dataset feasibility study. Verified datasets from Open Meteo, Sentinel or Landsat, CY-Bench Dataset for accurate crop monitoring and yield prediction.
Karanbir Singh	Engaged in a real-world farmer interview to understand practical issues in farming.
Sonu Choubey, Komal Dadwal, Jatin Mahey	Performed detailed literature review of current research papers to understand existing technologies, limitations, and how our project can provide innovation.
Mentor Guidance	Guided the team to refine focus initially planned to cover all AgriTech tracks (soil, irrigation, pest, monitoring, post-harvest), but later concentrated solely on **Crop Monitoring** for deeper innovation and better feasibility.

Approach to the Problem:

The project began with the problem of unreliable yield predictions and poor real-time crop health tracking by reading present problems in yield prediction machine learning models. A

data-driven approach was chosen, beginning with exploratory analysis of CY-Bench. Key challenges included merging spatial and tabular data and building interpretable models. After feature extraction and cleaning, models will be train and tune using k-fold validation.

Obstacles Faced:

- **Data Format Complexity**: CY-Bench includes mixed formats raster, CSV, shapefiles which required custom parsing scripts.
- **NDVI Area Constraint**: Processing large Sentinel or Landsat tiles efficiently for user-selected <5 km² regions needed optimization via GEE filters.

Literature Review:

S. No.	Author (First)	Year	Paper Title	Methodology Used	Gaps Found
1	Subramaniam et al.	2024	Crop yield prediction using effective DL and DR approaches for Indian regional crops	Preprocessing (cleaning/normalization), SEKPCA for dimensionality reduction, WTDCNN trained with Enhanced Whale Optimization Algorithm for yield prediction.	Incorporation of satellite imagery and remote sensing data for broader generalization.
2	Ashfaq et al.	2024	Accurate wheat yield prediction using ML and climate-NDVI data fusion	Used Google Earth Engine data (climate, soil, NDVI), tested SVM, RF, and LASSO. RF gave best results (RMSE = 0.05 q/ha, R^2 = 0.88).	Multi-source data integration (NDVI + weather + soil + spatial data) for local and regional predictions.
3	Jabed et al.	2024	Crop yield prediction in agriculture: A comprehensive review	Systematic Literature Review on ML/DL in CYP (2018–2023), guided by PRISMA. Analyzed features, models, evaluation metrics, and sustainability.	Lack of explainable AI (XAI), black-box model issues, limited data availability, and lack of multimodal data fusion.
4	Killeen et al.	2024	Corn grain yield prediction using UAV imagery, ML, and spatial CV	Used UAV RGB/MS images, evaluated 55 vegetation indices, trained RF and LR models with both standard and spatial CV. Found overfitting with 10-fold CV.	True spatial validation, generalization across locations and seasons using high-resolution imagery and NDVI. Didn't used feature engineering and hyperparameter tuning.
5	Sudhamathi et al.	2024	Ensemble regression based Extra Tree Regressor for hybrid crop yield prediction system	ER-ETR model using KPCA and LASSO; outperformed other models in MAE, MSE, RMSE, and F1-score.	Integration of prediction into decision support tools; overfitting concerns managed using

					generalizable, multimodal datasets.
6	Van Klompenburg et al.	2020	Crop yield prediction using machine learning: A systematic literature review	SLR on 50 ML papers and 30 DL-based studies. Identified top models (ANN, CNN, LSTM) and common challenges (data scarcity, model deployment gaps).	Focused farm-level application, integration into farm systems, and support for real-world decision making using diverse data.
7	Nikhil et al.	2024	ML-based crop yield prediction in South India: Performance analysis	Merged weather, soil, and crop data from public repositories. Trained multiple regression models, with Extra Trees Regressor achieving best R^2 = 0.9615.	Use of region-specific, real-world data from India; need for diverse and interpretable predictors in modelling.
8	Paudele et al.	2025	CY-Bench: A comprehensive benchmark dataset for sub-national crop yield forecasting	Developed a standardized multi-modal dataset for crop yield prediction with global coverage. Includes climate, remote sensing, soil, and socio-economic features.	First Indian project using CY-Bench for practical field-level forecasting with explainability, multimodal fusion, and decision-making.

How Our Project Fills These Gaps:

Our project systematically addresses the critical gaps in existing research as identified above:

- Multimodal Data Fusion: We integrate NDVI (via Sentinel or Landsat/Bhuvan NRSC), satellite-based soil/weather data (Open Meteo) to overcome single-source dependency.
- Explainability and Transparency: We apply Explainable AI (XAI) to provide transparency in our model, a gap strongly highlighted in multiple studies (e.g., Jabed, Van Klompenburg).
- CY-Bench Integration: We are the first in India to operationalize the CY-Bench dataset for real-world applications enhancing generalizability and reproducibility across regions and crops.

Additionally, our project addresses key methodological gaps in previous work like Killeen et al. by explicitly including feature engineering during data preprocessing and applying hyperparameter tuning both of which were either missing or not emphasized in their approaches.

References:

1. Subramaniam, Leelavathi Kandasamy, and Rajasenathipathi Marimuthu. "Crop yield prediction using effective deep learning and dimensionality reduction approaches for

Indian regional crops." *e-Prime–Advances in Electrical Engineering, Electronics and Energy* 8 (2024): 100611.

2. Ashfaq, Muhammad, et al. "Accurate wheat yield prediction using machine learning and climate-NDVI data fusion." *IEEE Access* 12 (2024): 40947-40961.
3. Jabed, Md Abu, and Masrah Azrifah Azmi Murad. "Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability." *Heliyon* (2024).
4. Killeen, Patrick, et al. "Corn grain yield prediction using UAV-based high spatiotemporal resolution imagery, machine learning, and spatial cross-validation." *Remote Sensing* 16.4 (2024): 683.
5. Sudhamathi, T., and K. Perumal. "Ensemble regression based Extra Tree Regressor for hybrid crop yield prediction system." *Measurement: Sensors* 35 (2024): 101277.
6. Van Klompenburg, Thomas, Ayalew Kassahun, and Cagatay Catal. "Crop yield prediction using machine learning: A systematic literature review." *Computers and Electronics in Agriculture* 177 (2020): 105709.
7. Nikhil, Uppugunduri Vijay, et al. "Machine learning-based crop yield prediction in south india: performance analysis of various models." *Computers* 13.6 (2024): 137.
8. Paudel, Dilli, et al. "CY-Bench: A comprehensive benchmark dataset for sub-national crop yield forecasting." *Earth System Science Data Discussions* (2025): 1-28.