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	Muhammad Ashfaq et al.	2024	Accurate Wheat Yield Prediction Using ML and Climate-NDVI Data Fusion	SVM, RF, LASSO on NDVI + weather + soil data via GEE	No feature engineering or hyperparameter tuning; limited explainability.
2	Khilola Amankulova et al.	2024	A Novel Fusion Method for Soybean Yield Prediction Using Sentinel-2 and PlanetScope Imagery	Focus on image classification-based predictions and remote sensing for irrigated lands.	Less relevant to crop yield prediction and lacks integration with ML and XAI.
3	Zeeshan Ramzan et al.	2023	Multimodal Data Fusion for Tea Yield Estimation Using Deep Neural Networks	Fusion of Landsat-8 NDVI and agromet data; NAS-based DNN architecture	Lacked soil, socio-economic, or fair generalization across regions; no yield expectation alignment.
4	Fudong Lin et al.	2023	MMST-ViT: Climate Change-aware Crop Yield via Vision Transformers	Multi-modal spatial-temporal ViT with Sentinel-2 + HRRR climate model data.	Model overfitting concerns; no real-time yield expectation or XAI explanation
5	Pascal Janetzky et al.	2024	Global Vegetation Modeling using Pre-trained Weather Transformers.	FourCastNet finetuning for global NDVI modeling from ERA5 data	Coarse temporal scale for extreme event analysis; lacks crop-specific context.
6	Dilli Paudel et al.	2023	Weakly Supervised Framework for High-Resolution Crop Yield Forecasts	Weak supervision using HR predictors and LR yield labels.	Absence of HR ground truth; fair for Europe but lacks Indian context.
7	Michiel Kallenberg et al.	2023	Process-based and ML Hybrid Yield Models	Meta-model: Tipstar crop model + CNN pre-trained on synthetic data	Process-based calibration complexity; lacks real-time features

					and fair comparison across datasets.
8	Dilli 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., Fudong Lin et al.)
- 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 Muhammad Ashfaq 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. Ashfaq, Muhammad, et al. "Accurate wheat yield prediction using machine learning and climate-NDVI data fusion." IEEE Access 12 (2024): 40947-40961.
2. Amankulova, Khilola, et al. "A Novel Fusion Method for Soybean Yield Prediction Using Sentinel-2 and PlanetScope Imagery." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2024).
3. Ramzan, Zeeshan, et al. "A multimodal data fusion and deep neural networks-based technique for tea yield estimation in Pakistan using satellite imagery." IEEE Access 11 (2023): 42578-42594.
4. Lin, Fudong, et al. "Mmst-vit: Climate change-aware crop yield prediction via multi-modal spatial-temporal vision transformer." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
5. Janetzky, Pascal, et al. "Global Vegetation Modeling with Pre-Trained Weather Transformers." ICLR 2024: Tackling Climate Change with Machine Learning: Fostering the Maturity of ML Applications for Climate Change. 2024.
6. Paudel, D. R., et al. "A weakly supervised framework for high-resolution crop yield forecasts." ICLR 2022: AI for Earth and Space Science. 2022.

7. Kallenberg, Michiel, et al. "Integrating processed-based models and machine learning for crop yield prediction." ICML 2023 Workshop: Synergy of Scientific and Machine Learning Modeling. 2023.
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.