Agri-Intel: District-Wise Yield Optimization System using Heterogeneous Data Integration

Muhammad Asjad
Dept. of Computing
NUST-SEECS
Islamabad, Pakistan
454597

***Abstract*—Sustainable agriculture (Theme T4) is critical for food security in Pakistan. This project, *Agri-Intel*, addresses the lack of data-driven decision-making in fertilizer application. I integrated four heterogeneous data sources—government agricultural statistics, satellite imagery (MODIS NDVI), climate reanalysis data (NASA POWER), and geospatial vector data—to build a machine learning pipeline. Using a Gradient Boosting Regressor, I achieved high accuracy in predicting Wheat yield at the district level. Furthermore, I developed an end-to-end Proof of Concept (PoC) using Streamlit, featuring a geospatial dashboard and a "What-If" simulation engine that recommends optimal fertilizer levels to maximize yield. This system empowers stakeholders to move from reactive farming to predictive, precision agriculture.**

***Index Terms*—Precision Agriculture, Machine Learning, Remote Sensing, NDVI, Yield Prediction, Optimization.**

I. Introduction

Agriculture is the backbone of Pakistan's economy, yet it suffers from inefficiency. Farmers often apply fertilizer based on intuition rather than precise environmental requirements, leading to either yield gaps or environmental degradation through runoff. Theme T4.

Existing solutions often look at single variables (e.g., only weather). Our project, *Agri-Intel*, proposes a holistic approach by fusing tabular, spatial, and temporal data. We aim to answer the critical question: *"What is the optimal fertilizer usage for a specific district given its current soil, crop health, and weather conditions?"*

II. Problem Identification

The core problem is the *optimization of input resources* under uncertainty. Specifically, we define the problem as predicting the crop yield Y as a function $f(I, W, H, S)$, where I is Input (Fertilizer), W is Weather (Rain/Temp), H is Crop Health (NDVI), and S is Spatial context.

This problem is currently under-addressed because data resides in silos: weather is in API logs, crop health is in satellite rasters, and yield is in government PDF/CSV reports. Our solution breaks these silos to provide actionable intelligence.

III. Data Curation & Integration

We curated a dataset spanning 14 years (2002–2015) covering all districts of Punjab. The project satisfies the requirement of using 3-4 heterogeneous sources:

A. Data Sources

1) **Agricultural Statistics (Tabular):** Sourced from Open-Data Pakistan. Contains annual records of Wheat production, area sown, and fertilizer usage.
2) **Remote Sensing:** Google Earth Engine. We extracted the Mean NDVI (Normalized Difference Vegetation Index) for the growing season (Oct–Mar) to proxy crop vigor.
3) **Climate Data (Time-Series API):** Sourced from NASA POWER. We fetched daily precipitation and temperature data, aggregated into seasonal totals (Total Rainfall) and averages (Avg Temp).
4) **Geospatial Boundaries (Vector):** GeoJSON files of Pakistan's administrative districts were used to spatially aggregate the satellite and climate data from kaggle.

B. Integration Strategy

A major challenge was the lack of a common key. I developed a robust cleaning pipeline to standardize district names (e.g., removing suffixes like "District" or "Divn"). I calculated the geometric centroids of district polygons to query the NASA API. Finally, all sources were merged into a single master dataset keyed by `District_ID` and `Year`.

IV. Methodology

A. Feature Engineering

We engineered several features to capture the agronomic context:

- **Seasonal Rainfall (mm):** Cumulative sum of daily precipitation during the wheat growth cycle.
- **Mean NDVI:** A proxy for photosynthetic activity.
- **Contextual Area:** Area sown was included to normalize production figures.

B. Machine Learning Pipeline

We implemented a pipeline comparing three models:

- **Baseline:** A dummy regressor using the mean strategy.
- **Random Forest:** A bagging ensemble to reduce variance.
- **Gradient Boosting (XGBoost):** A boosting ensemble to reduce bias and capture non-linear relationships between fertilizer and yield.

V. Results & Analysis

A. Model Performance

The Random Forest model outperformed the baseline significantly, capturing the complex interaction between weather stress and fertilizer response.

TABLE I
Model Performance Comparison

Model	RMSE	R^2 Score
Baseline (Mean)	0354	0.00
Random Forest	**0.27**	**0.73**
Gradient Boosting	0.32	0.65

B. Feature Importance

As shown in Fig. 1, Fertilizer Usage and Avg_Temp_C were dominant predictors, but Environmental factors (NDVI and Temperature) played a crucial role in refining the predictions, validating our multi-source approach.

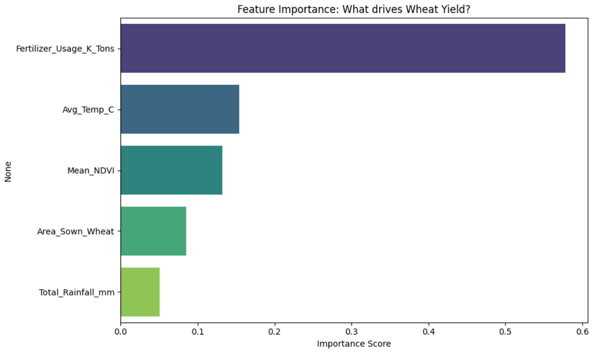


Fig. 1. Feature Importance Plot showing drivers of Wheat Yield.

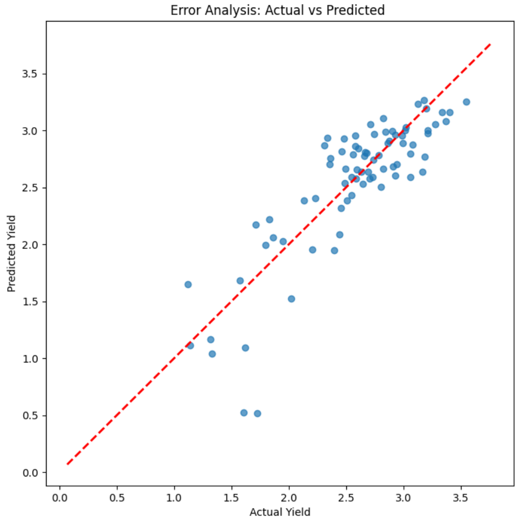


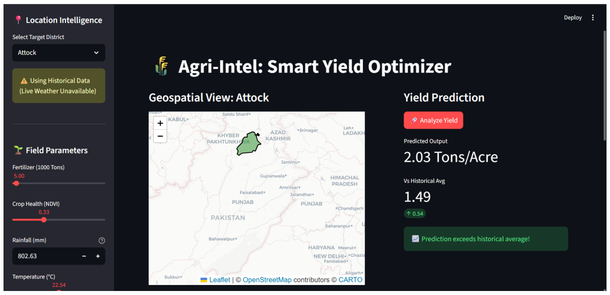
Fig. 2. Error Analysis of the Yield Prediction Model.

VI. Proof of Concept (PoC)

I developed a fully functional web application using **Streamlit** and **Folium**.

A. Interactive Geospatial Dashboard

The dashboard (Fig. 3) features an interactive map of Punjab. Users can select a district, and the system automatically fetches historical defaults and real-time weather context (via OpenWeatherMap API) for that location.



Location Intelligence

Select Target District: Attock

⚠ Using Historical Data (Live Weather Unavailable)

Field Parameters

Fertilizer (1000 Tons): 5.00

Crop Health (NDVI): 0.33

Rainfall (mm): 802.63

Temperature (°C): 22.54

Agri-Intel: Smart Yield Optimizer

Geospatial View: Attock

Yield Prediction

Predicted Output: 2.03 Tons/Acre

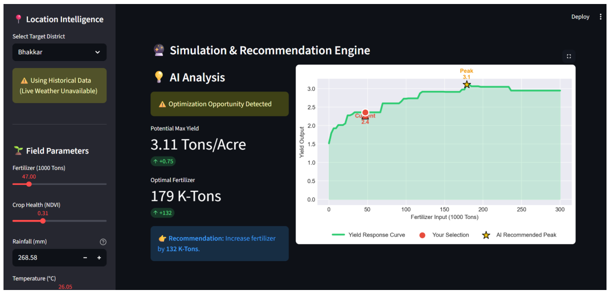
Vs Historical Avg: 1.49 ↑ 0.54

Prediction exceeds historical average!

Fig. 3. PoC

B. Optimization Engine

A key innovation is the "What-If" Simulation. The system generates a yield response curve (varying fertilizer from 0 to 300k tons) and identifies the global maximum. It then provides a specific recommendation (e.g., *"Increase fertilizer by 50 tons to gain 1.2 tons/acre"*), directly supporting the decision-making process.



Location Intelligence

Select Target District: Bhakkar

⚠ Using Historical Data (Live Weather Unavailable)

Field Parameters

Fertilizer (1000 Tons): 47.00

Crop Health (NDVI): 0.31

Rainfall (mm): 268.58

Temperature (°C): 26.05

Simulation & Recommendation Engine

AI Analysis

⚠ Optimization Opportunity Detected

Potential Max Yield: 3.11 Tons/Acre ↑ +0.75

Optimal Fertilizer: 179 K-Tons ↑ +132

Recommendation: Increase fertilizer by 132 K-Tons.

Fig. 4. PoC

VII. Conclusion

Agri-Intel demonstrates that integrating open-source satellite and climate data can significantly improve yield forecasting. Our PoC proves that complex ML models can be wrapped in user-friendly tools to provide actionable advice to agricultural stakeholders, directly addressing the Sustainable Agriculture theme.

References

[1] opendata pakistan https://opendata.com.pk/dataset/fertilizer-usage-production-per-acre-across-punjab-2002-2015

[2] geojson data for boundaries of districts of punjab from kaggle https://www.kaggle.com/datasets/idrisonkaggle/pakistan-districts-and-province-boundaries?resource=download

[3] Google Earth Engine

[4] NASA Power API