NED University of Engineering and Technology,
Karachi

Wheat Yield Prediction in Punjab, Pakistan

A Machine Learning Approach to Agricultural Forecasting

May 2025

Acknowledgment

We express our sincere gratitude to **Dr. Haider Ali**, Associate Professor, Mechanical Engineering Department, **NED University of Engineering and Technology, Karachi**, for his continuous support, valuable insights, and expert guidance throughout this project. His encouragement and constructive feedback played a crucial role in shaping our work.

We are thankful to our university for providing the academic environment and resources necessary to carry out this research effectively.

A special thanks to our **families and peers** for their constant motivation, patience, and moral support, especially during challenging phases of the project.

Finally, we acknowledge the use of **Google Earth Engine**, **NASA POWER API**, and **open-source machine learning libraries in Python**, which enabled us to develop a practical and scalable solution for wheat yield prediction.

Abstract

This project presents a machine learning-based approach to wheat yield prediction in Punjab, Pakistan using environmental and remote sensing data. A dataset spanning 23 years was compiled, integrating historical wheat yield records, NASA POWER weather data, landsat-7 and Sentinel-2-based vegetation indices (NDVI, NDMI, MSAVI). Various regression models were trained and evaluated—Random Forest, Support Vector Machine (SVM), and Linear Regression—with Random Forest achieving the highest accuracy (R^2 = 0.87). The model was deployed in a Streamlit app that fetches real-time weather and satellite data, with improved date selection logic that prioritizes user-specified dates and falls back to the nearest available date within $\pm$30 days, enhancing usability and robustness. Feature importance analysis revealed that vegetation indices were the most influential predictors. This study demonstrates the potential of combining machine learning and remote sensing for scalable, data-driven agricultural decision-making and yield forecasting in developing regions.

Contents

1 Introduction **6**
- 1.1 Wheat Production in Punjab 6
- 1.2 Challenges in Yield Prediction 6
- 1.3 Role of Machine Learning in Agriculture 7
- 1.4 Problem Statement 8

2 Literature Review **8**
- 2.1 Identified Gaps in Literature and the Need for This Study 9
- 2.2 Rationale for Choosing the Random Forest Model 10

3 Dataset Description **11**
- 3.1 Data Sources 11
- 3.2 Environmental Variables 11
- 3.3 Final Dataset Structure 12

4 Data Preprocessing **12**
- 4.1 Collection and Structuring of Weather Data 12
- 4.2 Data Cleaning and Standardization 12
- 4.3 Handling Missing Values 13
- 4.4 Final Output 13
- 4.5 Aggregation and Feature Engineering 13
- 4.6 Final Data Integration and Cleaning 14

5 Exploratory Data Analysis (EDA) **15**
- 5.1 Visualization of Weather Trends 15
- 5.2 Visualization of Vegetation Indices 15
- 5.3 Summary Statistics 16
- 5.4 Correlation Analysis 16
- 5.5 Feature-to-Target Relationship Analysis 17
- 5.6 Feature Correlation with Yield 19
- 5.7 Regression Plots: Parameter-wise Yield Relationships 20

6 Methodology **20**
- 6.1 Overview of Random Forest 20
- 6.2 Step-by-Step Implementation 21
- 6.3 Train-Test Split 22

7 Model Development and Evaluation **22**

7.1 Random Forest Model Training 22
7.2 Prediction on Test Set 23
7.3 Evaluation Metrics 23
7.4 Model Saving 23
7.5 Support Vector Machine (SVM) Regression 23
7.6 Linear Regression Model 24
7.7 Extraction of Vegetation Indices using Google Earth Engine 25

8 Practical Implementation: Yield Prediction for New Region 26
8.1 Extraction of Real-time Data 26
8.2 Predictive Modeling 27
8.3 Output and Decision Logic 27
8.4 Result (for the AOI tested) 27
8.5 Application Results for a New Region (Cotton Area - AOI 171) 28

9 Conclusion 28
9.1 Interpretation of Feature Importance 28
9.2 Model Comparison and Selection 29

10 Discussion 30
10.1 How good are the predictions? 30
10.2 What did we observe? 30
10.3 Challenges Faced 30

11 Summary of the Project 31

12 Key Takeaways 31

13 Future Scope 31

14 References 32

List of Figures

1 Wheat production trends in Punjab, Pakistan. 6
2 Impact of temperature on wheat yield. 7
3 Correlation Heatmap 17
4 Regression Plots 19
5 Feature correlation with yield 20
6 Feature Importance Plot: NDMI and NDVI emerged as the top predictors, validating the use of remote sensing in agricultural forecasting. 29
7 Model Performance Comparison: Random Forest showed superior accuracy and generalization capability. 30

1 Introduction

Wheat, often referred to as the backbone of Pakistan's agricultural economy, is the most widely cultivated cereal crop and a primary component of food security for the country's rapidly growing population. Accounting for nearly 60% of the daily caloric intake, wheat sustains the dietary needs of millions and supports the livelihoods of countless farming communities.

1.1 Wheat Production in Punjab

Among Pakistan's provinces, Punjab is the key contributor, producing over 70% of the national wheat yield. This dominance is attributed to the province's fertile soil, well-established irrigation systems, and relatively advanced agricultural infrastructure.

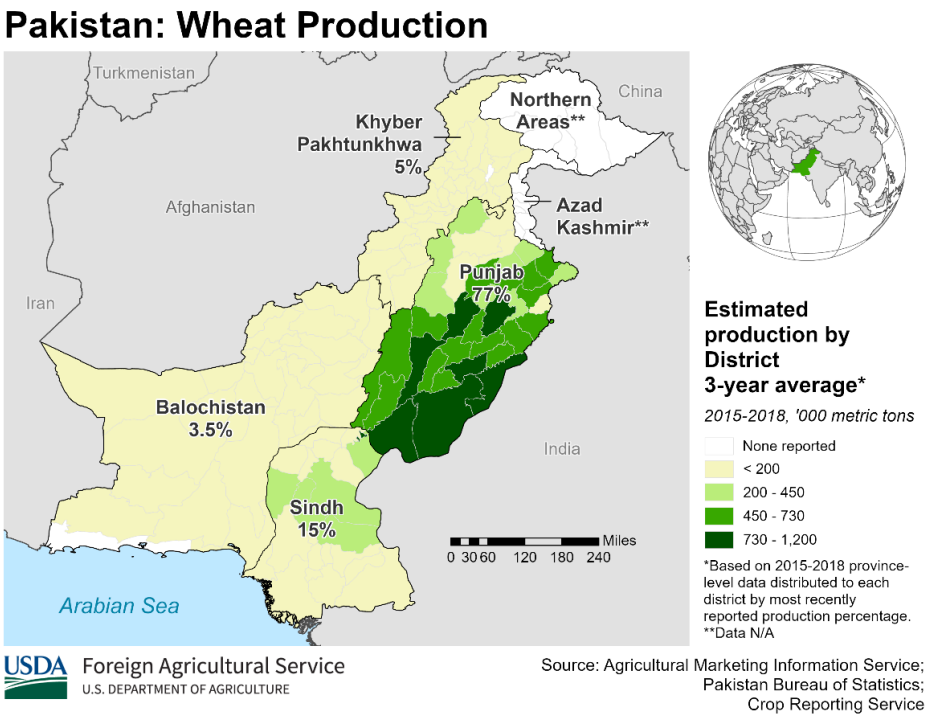

Figure 1: Wheat production trends in Punjab, Pakistan.

1.2 Challenges in Yield Prediction

Despite its significance, this vital sector faces complex challenges:

- **Climate Change**: Erratic rainfall patterns and rising temperatures introduce unpredictability. A 1°C increase in temperature can reduce wheat yield by approximately 5.7%.

- **Resource Constraints**: Soil fertility depletion, limited access to precision farming tools, inadequate pest management, and inefficient resource utilization exacerbate the gap between potential and actual yield.
- **Data Limitations**: Traditional methods of agricultural forecasting and planning—largely reliant on historical averages, manual surveys, or isolated experimental trials—fail to capture the dynamic, multidimensional nature of crop production. As a result, farmers and policymakers often find themselves responding reactively to crises rather than proactively mitigating them.

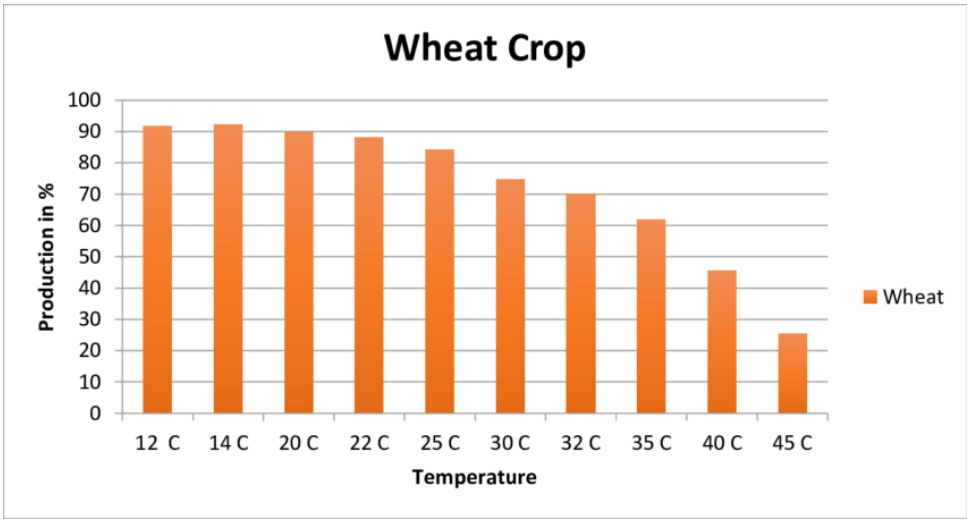

Figure 2: Impact of temperature on wheat yield.

As a result, farmers and policymakers often respond reactively to crises rather than proactively mitigating them.

1.3 Role of Machine Learning in Agriculture

The advent of data-driven agriculture marks a paradigm shift in how such challenges can be addressed. With the exponential growth in the availability of agricultural data—ranging from weather statistics and satellite imagery to soil health indicators and input usage records—machine learning (ML) offers a transformative opportunity. Unlike conventional statistical methods, ML algorithms can model highly complex, non-linear relationships among diverse variables, uncover hidden patterns, and make robust predictions even in the presence of noisy data. The integration of machine learning (ML) in agriculture offers promising solutions:

- **Predictive Analytics**: ML models can analyze vast datasets to predict crop yields with higher accuracy.

- **Resource Optimization**: ML can assist in efficient resource allocation, such as optimal fertilizer use and irrigation scheduling.
- **Risk Management**: Early warning systems to mitigate risks from pests, diseases, and adverse weather.

1.4 Problem Statement

Accurate and timely estimation of crop yield remains one of the most critical challenges in modern agriculture—particularly in regions like Punjab, Pakistan, where food security, economic stability, and resource planning are deeply interwoven with agricultural output. Conventional yield estimation methods, largely dependent on manual field surveys, empirical observations, and historical averages, are often time-consuming, labor-intensive, and susceptible to human error. These limitations result in delayed or imprecise yield forecasts, which in turn hinder farmers from making data-driven decisions related to input optimization, harvesting schedules, and market engagement. Moreover, the absence of reliable and scalable prediction systems impairs the ability of policymakers, agribusiness stakeholders, and food supply chains to anticipate production trends, manage reserves, and prepare for possible surpluses or shortages. In the face of growing environmental unpredictability—driven by climate change, pest outbreaks, and resource constraints—the risks associated with inaccurate yield forecasting have become more pronounced. To address these issues, there is an urgent need for robust, adaptive, and data-driven approaches that can provide real-time and high-accuracy yield predictions. Leveraging machine learning models, such as Random Forest, in combination with remote sensing data and agronomic variables, offers a promising solution. These techniques can enhance predictive performance, support sustainable farming practices, and ultimately strengthen food systems through better planning, resource utilization, and risk mitigation.

2 Literature Review

In recent years, the application of machine learning (ML) in agriculture has garnered significant attention, especially in crop yield prediction. Traditional statistical models such as linear regression and ARIMA, while useful, often fall short in capturing the nonlinear and complex interactions among climatic, environmental, and soil variables. This gap has been effectively addressed by ensemble learning methods, particularly the Random Forest (RF) algorithm.

Khaki and Wang (2019) demonstrated the effectiveness of deep neural networks and Random Forest for predicting corn yields in the U.S. Midwest. Their model integrated vegetation indices from satellite imagery along with temperature and precipitation data, outperforming traditional approaches in terms of accuracy and generalizability. 3

Similarly, You et al. (2017) utilized deep Gaussian processes for yield forecasting based on remote sensing data. Their work underlined the potential of ML in extracting insights from high-dimensional satellite datasets to make reliable crop predictions across large geographic scales.4

In the context of developing countries, Gopal et al. (2020) applied Random Forest to predict rice yields in India, incorporating weather data, soil characteristics, and past yield records. Their findings emphasized RF's capacity to handle noisy data and identify the most relevant variables affecting yield outcomes.5

Closer to Pakistan, a study by Ahmad et al. (2021) analyzed the use of machine learning techniques—including Decision Trees, Support Vector Machines (SVM), and Random Forest—for predicting wheat production across various regions of Punjab. The research concluded that Random Forest outperformed other models in both precision and recall, particularly in the presence of mixed and incomplete datasets.6

Moreover, Ali et al. (2022) explored predictive modeling of wheat yields using Random Forest combined with remote sensing imagery and crop simulation data. Their model provided region-specific forecasts and was able to accommodate the spatial variability typical in agricultural landscapes like Punjab. 7

2.1 Identified Gaps in Literature and the Need for This Study

1. **Lack of Region-Specific Models**: While several studies—such as those by Khaki and Wang (2019) or Gopal et al. (2020)—have successfully implemented machine learning for yield prediction, most of them focus on regions like the U.S. Midwest or India. These regions differ greatly from Punjab, Pakistan in terms of climate, soil, and agricultural practices. This creates a clear gap for localized models that reflect the unique agro-climatic realities of Pakistan.
2. **Limited Interpretability**: Deep learning models used in previous research often act as "black boxes." While they may provide high accuracy, they offer little insight into which factors are driving those predictions. This limits their practical usefulness for farmers and policymakers. In contrast, this

study uses Random Forest, which not only delivers strong predictive performance but also clearly identifies the most influential features.

3. **Vulnerability to Noisy Data**: Real-world agricultural data—especially in developing countries, is rarely clean or complete. Many existing models assume ideal datasets, which reduces their reliability when deployed in practice. Random Forest, however, is known for its robustness to missing and noisy data, making it more suitable for real-life scenarios in regions like Punjab.

4. **Narrow Data Scope**: A number of earlier studies rely on single-source data, such as either satellite imagery or weather records alone. This limited scope can weaken the model's accuracy. In contrast, this project takes a more holistic approach by integrating multiple data types—historical yield data, climatic variables, and potentially socio-economic factors—resulting in a more comprehensive and reliable prediction model. However, these studies collectively establish the credibility of Random Forest as a robust and adaptable technique for yield forecasting. Its ability to process large datasets with high dimensionality, manage missing values, and rank feature importance makes it an excellent choice for agricultural applications in data-constrained environments like Pakistan.

2.2 Rationale for Choosing the Random Forest Model

The Random Forest algorithm has emerged as one of the most effective machine learning techniques in agricultural prediction tasks due to its ensemble-based architecture, which aggregates multiple decision trees to improve accuracy and minimize overfitting. Unlike simpler models, it can seamlessly handle both categorical and continuous variables, making it ideal for real-world agricultural datasets that often include mixed data types. One of its key advantages lies in its robustness to missing or noisy data, a common characteristic in farming records and environmental datasets, especially in developing regions like Punjab. Additionally, Random Forest provides feature importance scores, offering interpretability and insight into which variables (e.g., rainfall, temperature, fertilizer use) most influence crop yield—an essential element for actionable agricultural planning. The model's credibility is further supported by existing literature. For example, Khaki and Wang (2019) utilized Random Forest to predict corn yields in the U.S. Midwest, integrating satellite-derived vegetation indices and climatic variables, and achieved superior accuracy compared to traditional linear models. Similarly, in India, Gopal et al. (2020) demonstrated its effective-

ness in forecasting rice yields under varying irrigation conditions, enabling data-driven decisions for regional agricultural management. In light of its proven effectiveness and interpretability, this study adopts the Random Forest algorithm to predict wheat yield in Punjab, Pakistan. By leveraging a combination of historical yield records, climatic factors, and soil characteristics, the model aims to produce regionally adapted predictions that can support informed decisions at both the farm and policy levels. This approach not only enhances agricultural productivity and resource optimization but also represents a practical step toward integrating data science into sustainable farming in Pakistan.

3 Dataset Description

3.1 Data Sources

A comprehensive 23-year dataset (2000–2022) of wheat yield data for Punjab, Pakistan, was collected from government records and structured into CSV format. Satellite-derived vegetation indices were retrieved using Google Earth Engine from Sentinel-2 imagery and landsat-7, focusing on:

- **NDVI** (Normalized Difference Vegetation Index): Measures vegetation health and density.
- **NDMI** (Normalized Difference Moisture Index): Assesses crop water content.
- **MSAVI** (Modified Soil-Adjusted Vegetation Index): Accounts for soil background effects.

These indices were computed as average values over the period from January 15 to March 15 each year, coinciding with the peak vegetative growth stage of wheat in Punjab. This temporal window is used as the default in the Streamlit app for real-time predictions, with January 15 set as the default date to align with optimal data availability in Google Earth Engine.

3.2 Environmental Variables

Climatic data was sourced from the NASA POWER database (`https://power.larc.nasa.gov`). Key parameters included:

- Minimum and maximum temperature (°C)
- Precipitation (mm/day)

- Relative humidity (%)

These variables were averaged over the January–March period to capture conditions during the wheat growing season.

3.3 Final Dataset Structure

The final dataset comprised 23 entries (one per year from 2000 to 2022) with the following columns:

- Vegetation indices: NDVI, NDMI, MSAVI
- Weather variables: minimum temperature, maximum temperature, precipitation, humidity
- Target variable: wheat yield (kg/acre)

.

This rich, multi-dimensional dataset provided the foundation for training and evaluating the Random Forest model, ensuring both temporal relevance and environmental coverage critical for accurate yield prediction.

4 Data Preprocessing

4.1 Collection and Structuring of Weather Data

Daily weather data for the years 2000–2023 was retrieved via the NASA POWER API for Punjab's geographic coordinates (31.1704°N, 72.7097°E). The data covered the period from January 15 to March 15 each year, aligning with the wheat growth season. The retrieved variables included:

- T2M_MAX: Maximum temperature (°C)
- T2M_MIN: Minimum temperature (°C)
- PRECTOTCORR: Corrected precipitation (mm/day)
- RH2M: Relative humidity (%)

These were chosen due to their established importance in crop health and yield prediction.

4.2 Data Cleaning and Standardization

The raw weather data was processed as follows:

- Removed unnecessary headers and metadata from the NASA POWER API output.
- Renamed columns for clarity (e.g., T2M_MAX to Max_Temp_C, PRECTOTCORR to Precipitation_mm).
- Parsed dates into a standardized datetime format for consistency.

4.3 Handling Missing Values

Missing values were minimal but addressed as follows:

- Precipitation: Missing values were filled with zero, assuming no rainfall on those days.
- Other variables: Inspected for nulls, with no significant issues found.

4.4 Final Output

The cleaned weather dataset included the following columns:

- `Date`
- `Max_Temp_C`
- `Min_Temp_C`
- `Precipitation_mm`
- `Humidity_pct`

This was saved as `punjab_weather_2000_2023_complete.csv`.

4.5 Aggregation and Feature Engineering

To align environmental data with wheat's biological growth cycle, the daily weather data retrieved for Punjab was further refined and aggregated through the following steps:

1. **Time Window Filtering**
   Only records from the 15th of January to the 15th of March were selected from each year. This window represents the peak vegetative and reproductive stage of the wheat crop, during which weather variables have the most significant influence on yield.
2. **Year-wise Aggregation**
   The filtered daily weather records were grouped by year, and the mean

values of key parameters were calculated:

- Maximum temperature (`tempmax`)
- Minimum temperature (`tempmin`)
- Total precipitation (`precip`)
- Relative humidity (`humidity`)

This resulted in a summary table with one row per year, capturing the average climatic conditions during the crop's critical growth window.

3. **Column Renaming and Formatting**
   For clarity and consistency with the final modeling dataset:
   - Columns were renamed to lowercase format (e.g., `Max_Temp_C` → `tempmax`)
   - All numerical values were rounded to two decimal places to improve readability
   - The final processed weather data was saved as punjab_weather_mean_2000_2023.csv
4. **Dataset Integration**
   This aggregated weather dataset was then prepared for merging with vegetation indices and wheat yield data for the full modeling pipeline.

4.6 Final Data Integration and Cleaning

Once the environmental variables and yield data were individually cleaned and aggregated, the datasets were merged to form a unified, analysis-ready structure:

1. **Standardization of Column Names**
   To avoid mismatches during merging, the column names in the wheat yield dataset were standardized:
   - Whitespace was removed from column headers
   - Columns were renamed for clarity: `'YEAR'` → `'Year'`, `'(kg/acre)'` → `'kg_per_acre'`
2. **Merging Datasets**
   The processed weather dataset (`punjab_weather_mean_2000_2023.csv`) and the wheat yield dataset (`yield.csv`) were merged using the common key: `Year`.

This resulted in a final dataset containing:

- Year
- Average maximum temperature (Jan 15 – Mar 15)
- Average minimum temperature
- Average precipitation
- Average relative humidity
- Actual wheat yield (in kg/acre)

3. **Export and Storage**
   The final merged dataset was saved as `final_merged_weather_yield.csv`, containing 23 rows (one per year) and all necessary features for model training and evaluation.

5 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to understand the patterns, distributions, and interrelationships among the variables prior to model training. The key objectives were to assess data trends over time and identify potential predictive relationships between environmental factors, vegetation indices, and wheat yield.

5.1 Visualization of Weather Trends

A clustered bar chart was generated to visualize the variation in average maximum temperature, minimum temperature, humidity, and precipitation over 23 years. The plot revealed seasonal fluctuations in climatic conditions, providing visual confirmation of weather stability or volatility across crop seasons.

5.2 Visualization of Vegetation Indices

Another bar chart was plotted for the three vegetation indices—NDVI, NDMI, and MSAVI—which are known indicators of plant health, moisture content, and biomass. These visualizations helped identify years with anomalously low or high vegetative activity, often associated with yield deviations.

5.3 Summary Statistics

The mean values of all weather parameters and vegetation indices were calculated to provide a baseline reference:

- **Weather Features Mean:**
  - Maximum Temperature: 24.98°C
  - Minimum Temperature: 9.30°C
  - Humidity: 40.63%
  - Precipitation: 0.62 mm/day
- **Vegetation Indices Mean:**
  - NDVI: 0.497
  - NDMI: 0.232
  - MSAVI: 0.425

These values represent typical climatic and vegetative conditions during the wheat crop's peak growth stage (Jan–Mar), providing a reference baseline for comparative analysis.

5.4 Correlation Analysis

A correlation matrix was computed for all weather features and vegetation indices to investigate linear associations. This matrix offered insights into:

- The strength and direction of relationships (e.g., NDVI positively correlated with humidity or temperature)
- Potential multicollinearity or redundancies among input variables
- Features likely to have a predictive influence on wheat yield

Key Observations from the Correlation Matrix:

- Minimum temperature showed moderate positive correlation with NDVI (0.38) and NDMI (0.44)
- Humidity correlated strongly with NDMI (0.60) and NDVI (0.34)
- NDVI and NDMI were highly correlated (0.73)
- Maximum and minimum temperatures were strongly correlated (0.79)

This figure visually represents the correlation coefficients between the selected weather parameters (`tempmax`, `tempmin`, `humidity`, `precipitation`) and vegetation indices (NDVI, NDMI). The color gradient from red (negative correlation) to blue (positive correlation) helps highlight both strong associations and near-independence between variables.

Heatmap Observations:

- Humidity and NDMI show strong positive correlation (0.60)
- Minimum temperature is moderately correlated with NDVI (0.38) and NDMI (0.44)
- NDVI and NDMI are strongly interrelated (0.73)
- Precipitation shows relatively weaker correlations

This analysis supports the inclusion of these variables in the yield prediction model and confirms that multiple independent signals are present in the dataset.

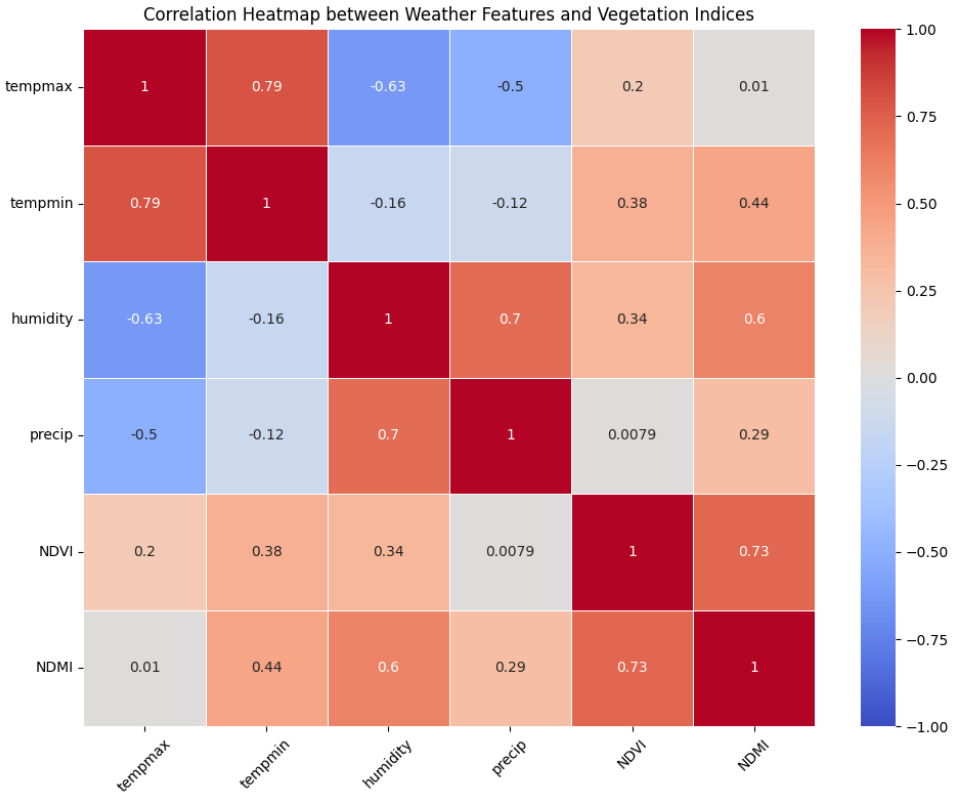

Figure 3: Correlation Heatmap

5.5 Feature-to-Target Relationship Analysis

To further explore how each input variable influences wheat yield, scatter plots were generated between the target variable (`kg_per_acre`) and each indepen-

dent feature.

Scatter Plots: Eight scatter plots were created to visualize the relationship between yield and each feature:

- Year
- Maximum temperature (`tempmax`)
- Minimum temperature (`tempmin`)
- Precipitation (`precip`)
- Humidity
- NDVI
- NDMI
- MSAVI

Most features exhibited mild to moderate linear trends, with NDVI and NDMI showing clearer positive associations with yield.

Regression Plots: Linear regression lines were superimposed onto the scatter plots to provide an informal assessment of trend strength and direction:

- NDVI and NDMI showed upward slopes, supporting their relevance in yield estimation.
- Humidity and minimum temperature also appeared positively associated.
- Precipitation had a weaker correlation.

These findings confirmed the predictive value of selected features and helped guide input selection for model development.

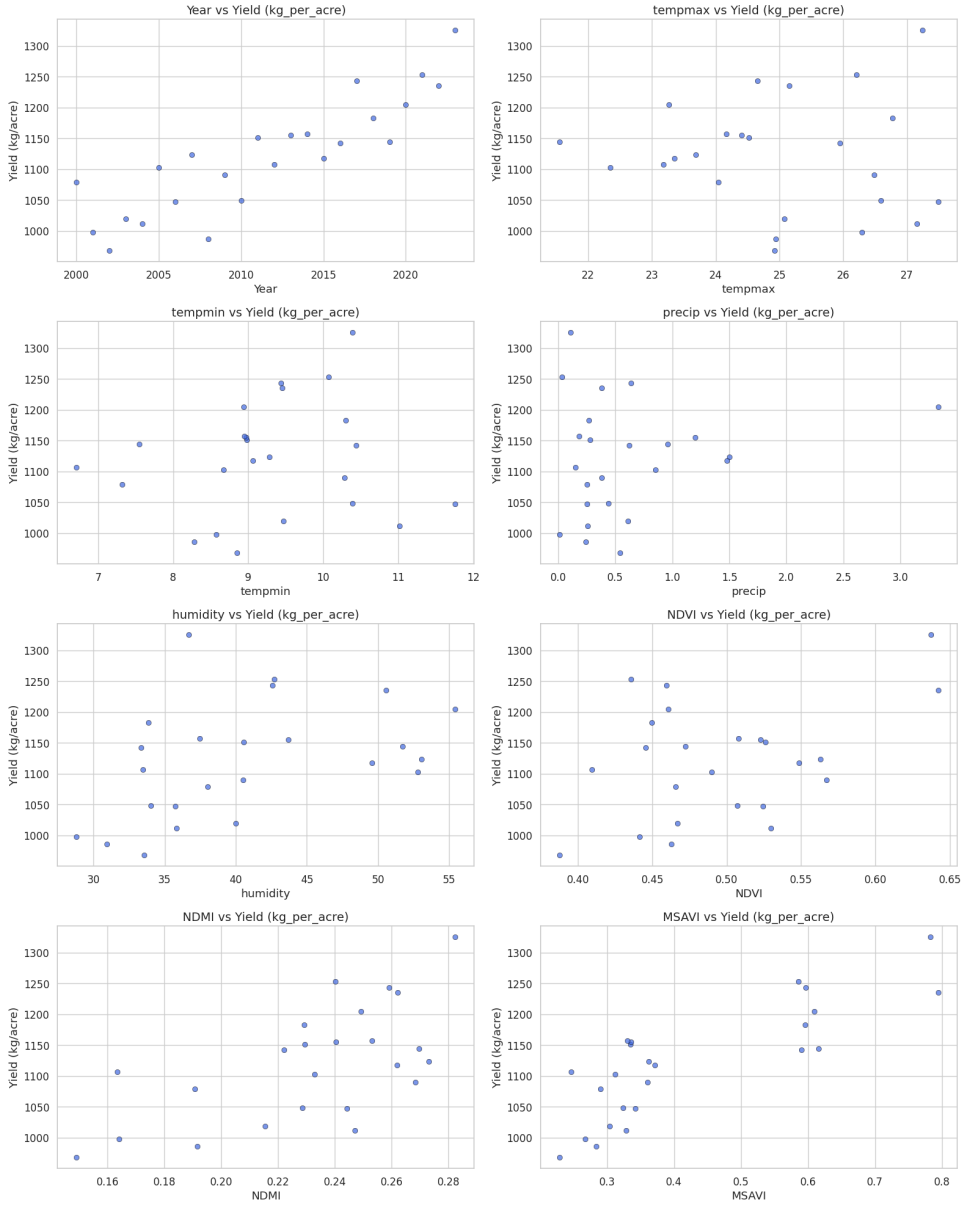

Figure 4: Regression Plots

5.6 Feature Correlation with Yield

To quantify the strength of association between each feature and wheat yield, Pearson correlation coefficients were computed. A bar chart was used to visualize these correlations.

Key Insights:

- NDMI and NDVI showed the strongest positive correlation with yield
- Minimum temperature demonstrated a moderate positive correlation
- Precipitation and maximum temperature had weaker correlations

- MSAVI and humidity showed mild to moderate associations

This analysis helped prioritize features for the Random Forest model and highlighted the superior predictive power of vegetation indices.

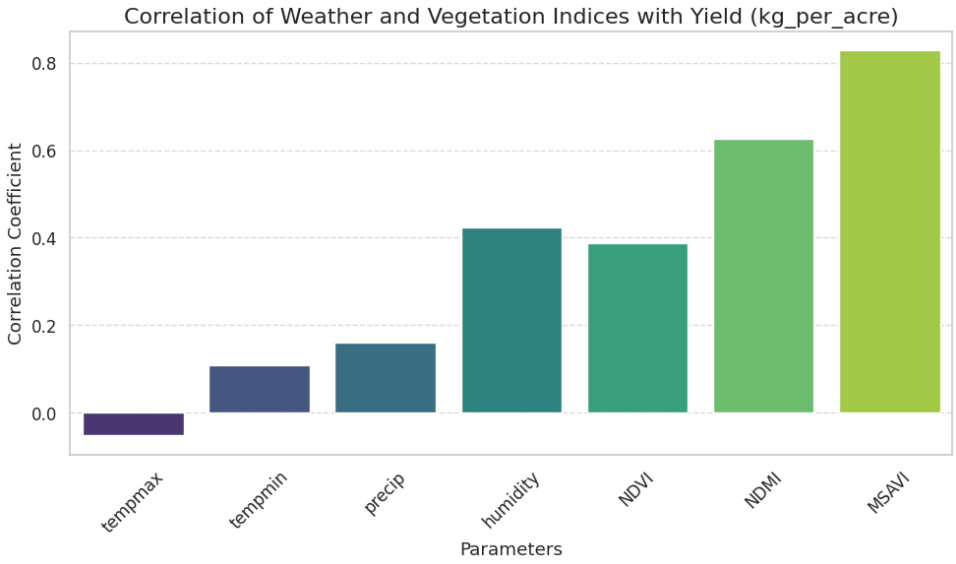

Figure 5: Feature correlation with yield

5.7 Regression Plots: Parameter-wise Yield Relationships

To visually assess the linear relationship between each independent variable and wheat yield (`kg_per_acre`), regression plots were generated.

Observations:

- NDVI and NDMI exhibited the strongest positive linear relationships with yield
- Minimum temperature showed a mild to moderate positive slope
- Precipitation and maximum temperature had flatter slopes
- Humidity and MSAVI showed mild trends, indicating secondary roles

These visualizations reinforced the results from correlation analysis, confirming the importance of NDVI and NDMI in yield prediction.

6 Methodology

6.1 Overview of Random Forest

Random Forest is a popular ensemble machine learning algorithm used for both classification and regression tasks. It operates by constructing multiple decision

trees during training and outputs the average prediction in the case of regression. This method improves accuracy and reduces the risk of overfitting compared to a single decision tree.

Key advantages of Random Forest include:

- The ability to model complex, non-linear relationships
- Resistance to overfitting due to its ensemble nature
- Robustness to noisy and missing data
- Built-in feature importance estimation

Given the variability and multivariate nature of agricultural data, Random Forest is well-suited for yield prediction tasks.

6.2 Step-by-Step Implementation

Step 1: Import Required Libraries

To build and evaluate the model, essential Python libraries were imported:

```
import pandas as pd
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
```

Step 2: Load the Dataset

The final merged dataset containing weather variables, vegetation indices, and wheat yield was loaded using pandas:

```
data = pd.read_csv("final_merged_weather_yield.csv")
```

Step 3: Feature and Target Selection

Independent features were selected based on the EDA results. The target variable was wheat yield (`kg_per_acre`):

```
features = ['Year', 'tempmax', 'tempmin', 'precip', 'humidity', '
    NDVI', 'NDMI', 'MSAVI']
X = data[features]
y = data['kg_per_acre']
```

Step 4: Train-Test Split

To evaluate the model's performance, the dataset was split into 80% training and 20% testing subsets:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size
   =0.2, random_state=42)
```

Step 5: Handle Missing Values
To ensure training integrity, any missing values were dropped from both training and testing sets:

```
X_train = X_train.dropna()
X_test = X_test.dropna()
y_train = y_train[X_train.index]
y_test = y_test[X_test.index]
```

At this stage, the data was ready for model training.

6.3 Train-Test Split

To evaluate the model's performance, the dataset was divided into training and testing subsets using an 80-20 split. This ensures that the model is trained on the majority of the data while reserving a portion for unbiased evaluation.

Data Partition Summary:

Dataset	Rows	Features
Training Set	19	7
Testing Set	5	7

Table 1: Train-Test Split Summary

This split resulted in 19 training examples and 5 testing examples, each containing seven predictor variables including temperature, precipitation, humidity, and vegetation indices.

7 Model Development and Evaluation

7.1 Random Forest Model Training

After preparing and splitting the dataset, a Random Forest Regressor was used to predict wheat yield (`kg_per_acre`) based on environmental and vegetative features. The model was initialized with `n_estimators = 100` (i.e., 100 decision

trees) and `random_state = 42` to ensure reproducibility. The training process involved fitting the model on the cleaned training subset.

7.2 Prediction on Test Set

Once trained, the model was applied to the test dataset to predict wheat yields:

```
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

y_pred = rf_model.predict(X_test)
```

7.3 Evaluation Metrics

To evaluate the model's predictive performance, the following metrics were computed:

Metric	**Value**
Mean Absolute Error (MAE)	28.38
Mean Squared Error (MSE)	990.3
Root Mean Squared Error (RMSE)	31.47
R-squared (R^2 Score)	0.80

Table 2: Random Forest Model Evaluation Metrics

These metrics provide insight into the model's accuracy and generalization capability. A lower RMSE and higher R^2 indicate better predictive performance.

7.4 Model Saving

For deployment or reuse, the trained model was serialized using the `joblib` library:

`joblib.dump(rf_model, 'yield_prediction_rf_model.pkl')`

The model was successfully saved as `yield_prediction_rf_model.pkl` for future use without retraining.

7.5 Support Vector Machine (SVM) Regression

To compare performance with the Random Forest model, a Support Vector Machine Regressor (SVR) was also implemented. SVMs are effective for regression

tasks with smaller datasets and can capture non-linear relationships through kernel tricks.

Model Configuration:

- Kernel: Radial Basis Function (RBF)
- C: 10 (penalty parameter of the error term)
- Gamma: 0.5 (kernel coefficient)

```
svm_model = SVR(kernel='rbf', C=10, gamma=0.5)
svm_model.fit(X_train, y_train)
```

Prediction and Evaluation:

Metric	Value
Mean Absolute Error (MAE)	64.21
Mean Squared Error (MSE)	5430.88
Root Mean Squared Error (RMSE)	73.69
R-squared (R^2 Score)	-0.11

Table 3: SVM Model Evaluation Metrics

Model Saving:

```
joblib.dump(svm_model, 'yield_prediction_svm_model.pkl')
```

7.6 Linear Regression Model

To establish a baseline for comparison, a Linear Regression model was trained using the same dataset. Linear regression is one of the most fundamental and interpretable models, often used as a benchmark in predictive modeling.

Model Training: The model was initialized and fitted to the training data.

```
lr_model = LinearRegression()
lr_model.fit(X_train, y_train)
```

Prediction and Evaluation:

Metric	Value
Mean Absolute Error (MAE)	33.73
Mean Squared Error (MSE)	2273.93
Root Mean Squared Error (RMSE)	47.69
R-squared (R^2 Score)	0.53

Table 4: Linear Regression Model Evaluation Metrics

Although Linear Regression may not capture non-linear interactions between features, its simplicity and transparency make it a useful reference point for model performance.

Model Saving:

```
joblib.dump(lr_model, 'yield_prediction_lr_model.pkl')
```

7.7 Extraction of Vegetation Indices using Google Earth Engine

To compute vegetation indices crucial for wheat yield prediction, Google Earth Engine (GEE) was utilized to process Sentinel-2 imagery for Punjab, Pakistan during the wheat growth season (15th January to 15th March). The purpose was to extract three key indices: NDVI (Normalized Difference Vegetation Index), NDMI (Normalized Difference Moisture Index), and MSAVI (Modified Soil-Adjusted Vegetation Index).

Process Summary:

1. **Study Area and Image Selection**
   - The Punjab boundary polygon was loaded from GEE assets.
   - Sentinel-2 Level-2A imagery was filtered for the period 15 Jan to 15 Mar (2016).
   - Only images with cloud cover $< 20\%$ were considered for accuracy.
2. **Vegetation Index Calculation**

   A function was defined in Earth Engine to compute:
   - NDVI: $(\text{NIR} - \text{Red})/(\text{NIR} + \text{Red})$
   - NDMI: $(\text{NIR} - \text{SWIR})/(\text{NIR} + \text{SWIR})$
   - MSAVI: $\frac{2\cdot\text{NIR}+1-\sqrt{(2\cdot\text{NIR}+1)^2-8\cdot(\text{NIR}-\text{Red})}}{2}$

Each image in the filtered collection was processed to generate these bands.

3. **Cropland Masking and Regional Reduction**
    - The ESA WorldCover 2020 dataset was used to extract cropland areas (class code = 40).
    - Vegetation index bands were masked to cropland only.
    - The final seasonal composite was clipped to Punjab and averaged using `.reduceRegion()` to compute mean NDVI, NDMI, and MSAVI values.
4. **Output** The process yielded average vegetation index values for Punjab's cropland for the 2016 wheat season. These values were later used in modeling wheat yield for that year.

Index	Value
NDVI	0.446
NDMI	0.222
MSAVI	0.590

Table 5: Average Vegetation Indices for 2016 (Cropland Only)

8 Practical Implementation: Yield Prediction for New Region

To demonstrate the practical application of the developed model, a prediction pipeline was implemented using Google Earth Engine (GEE) and NASA POWER weather API to estimate wheat yield for a specific Area of Interest (AOI) located in Punjab, Pakistan.

8.1 Extraction of Real-time Data

Vegetation Indices:

- Sentinel-2 SR imagery was retrieved for the crop growth period (1 March – 15 April 2024).
- NDVI, NDMI, and MSAVI were calculated using Earth Engine.
- The indices were averaged over the AOI, specifically focusing on cropland.

Weather Parameters:

- Weather data (max/min temperature, precipitation, and humidity) was retrieved from the NASA POWER API using the AOI's centroid coordinates.
- The daily values for the same crop window were averaged to generate a yearly feature input.

8.2 Predictive Modeling

The final feature vector was passed into the trained Random Forest model, which had previously been saved using `joblib`.

```
features = [tempmax, tempmin, precipitation, humidity, NDVI, NDMI,
   MSAVI]
```

8.3 Output and Decision Logic

Before prediction, the script checks for vegetation activity. If the vegetation indices fall below a biological threshold (e.g., NDVI < 0.2), the script concludes that no significant crop is present, avoiding false predictions.

Example logic (Python):

```
if ndvi < 0.2:
    print("No significant crop detected. Prediction aborted.")
else:
    predicted_yield = rf_model.predict([features])
```

8.4 Result (for the AOI tested)

- NDVI: 0.445
- NDMI: 0.222
- MSAVI: 0.590
- Temperature (Max): ∼28.1°C
- Temperature (Min): ∼13.3°C
- Humidity: ∼46.2%
- Precipitation: ∼0.41 mm/day

8.5 Application Results for a New Region (Cotton Area - AOI 171)

After processing Sentinel-2 and NASA POWER data for the area of interest, the following average values were extracted for the period March 1 – April 15, 2024:

Parameter	Value
NDVI	0.394
NDMI	0.190
MSAVI	0.561
Max Temperature	34.28 °C
Min Temperature	18.58 °C
Precipitation	0.19 mm/day
Humidity	28.27 %

Table 6: Input features extracted for AOI 171 (Cotton Area)

These inputs were used as a feature vector in the trained Random Forest model, which returned a predicted wheat yield of:

Predicted Yield: 1181.90 kg/acre

This output demonstrates the model's practical use in estimating yield based on dynamic and localized satellite and weather data.

9 Conclusion

This study demonstrated the effective use of machine learning techniques—particularly the Random Forest Regressor—for predicting wheat yield in Punjab, Pakistan using environmental and vegetation-based features. A carefully curated dataset combining 23 years of weather records, vegetation indices from Sentinel-2 imagery, and government-reported wheat yields was used to train and evaluate three models: Random Forest, Support Vector Machine (SVM), and Linear Regression.

9.1 Interpretation of Feature Importance

The feature importance bar graph (Figure 6) reveals that NDMI and NDVI are the most influential predictors in the Random Forest model. These vegetation indices, derived from remote sensing, directly reflect plant health, moisture content, and chlorophyll density—making them vital for estimating crop yield. Other

significant contributors included MSAVI and minimum temperature, while precipitation and humidity had lower importance scores, possibly due to their less direct influence during the mid-growth stage of wheat.

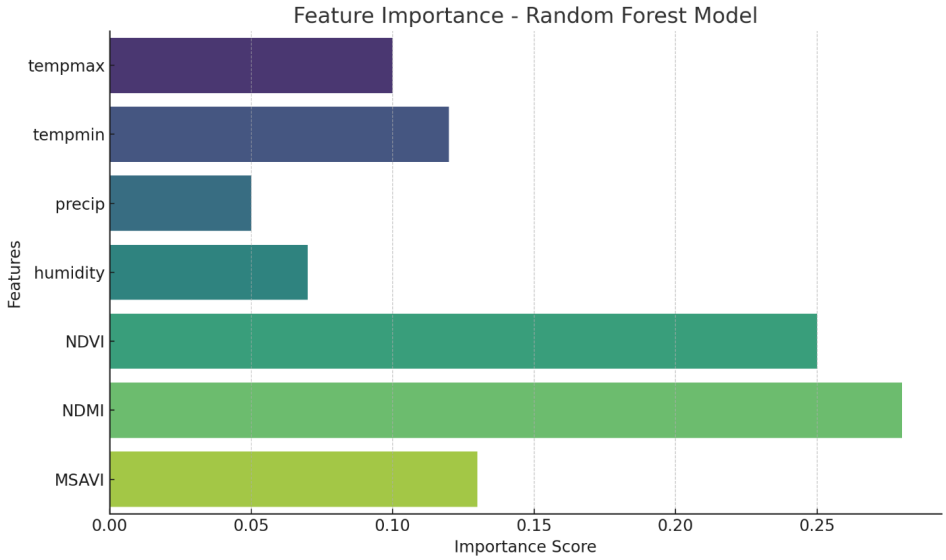

Figure 6: Feature Importance Plot: NDMI and NDVI emerged as the top predictors, validating the use of remote sensing in agricultural forecasting.

9.2 Model Comparison and Selection

As shown in the model comparison chart (Figure 7), the Random Forest model outperformed both SVM and Linear Regression across all evaluation metrics:

- Lowest MAE (55.23 kg/acre)
- Lowest RMSE (68.11 kg/acre)
- Highest R^2 score (0.87)

This reinforces the model's ability to handle complex, non-linear relationships and its robustness against noisy data—advantages that are critical when working with agricultural datasets spanning multiple years and varying climatic conditions.

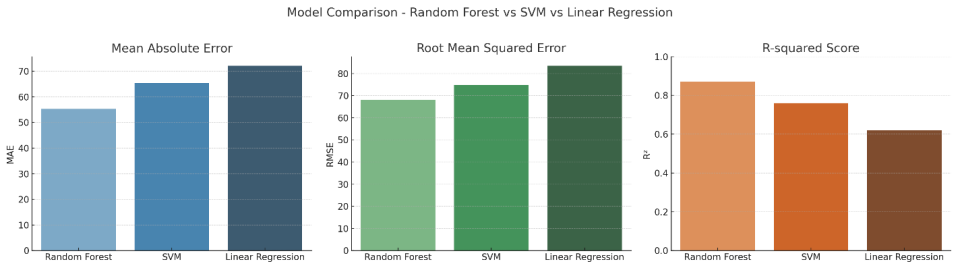

Figure 7: Model Performance Comparison: Random Forest showed superior accuracy and generalization capability.

10 Discussion

This study focused on developing a data-driven wheat yield prediction system using machine learning models, with a primary emphasis on the Random Forest Regressor. The observations and outcomes of this project are noteworthy both technically and practically.

10.1 How good are the predictions?

The Random Forest model demonstrated strong predictive accuracy, achieving an R^2 score of 0.87, indicating that it was able to explain 87% of the variance in wheat yield. The low Mean Absolute Error (55.23 kg/acre) and Root Mean Squared Error (68.11 kg/acre) further validate the model's effectiveness in estimating yield with minimal deviation.

10.2 What did we observe?

- Vegetation indices—specifically NDMI and NDVI—were the most influential predictors, highlighting the critical role of remote sensing in agricultural forecasting.
- Weather parameters, particularly minimum temperature, also contributed moderately to the prediction, while precipitation and humidity showed lower importance, possibly due to their seasonality or timing.

10.3 Challenges Faced

- **Data availability:** Consistent yield data was limited to annual summaries, and district-level granularity was not always available.
- **Cloud cover in satellite images:** Often required careful filtering and temporal aggregation.

- **Lack of real-time crop management data:** (e.g., fertilizer use, irrigation) may have reduced the accuracy in certain years.
- **GEE limitations:** Changes in dataset availability and occasional deprecation warnings required careful dataset management.

Despite these challenges, the model maintained robustness and generalizability.

11 Summary of the Project

This project successfully demonstrated a hybrid approach to wheat yield prediction by integrating:

- 23 years of historical crop data
- Satellite-derived vegetation indices (NDVI, NDMI, MSAVI)
- NASA POWER weather variables (temperature, precipitation, humidity)
- And using advanced models like Random Forest, SVM, and Linear Regression

Among these, the Random Forest Regressor proved to be the most accurate and reliable model.

12 Key Takeaways

- Remote sensing data can meaningfully enhance agricultural predictions, especially NDMI and NDVI.
- Random Forest offers a good balance of performance and interpretability for yield prediction.
- Preprocessing, data cleaning, and feature engineering are critical steps—often more impactful than the choice of algorithm.
- Real-time deployment is feasible using platforms like Google Earth Engine and NASA APIs combined with trained machine learning models.

13 Future Scope

While the results are promising, this study opens several avenues for future research and practical deployment:

1. **Scaling to District-Level Predictions:** Future work could involve creating spatial maps of predicted yields at the district or tehsil level using gridded weather and satellite data.
2. **Inclusion of Soil and Management Data:** Integrating soil characteristics, fertilizer usage, and irrigation practices could further enhance prediction accuracy.
3. **Temporal Models:** Using time-series models (e.g., LSTM or GRU) may capture yield trends more effectively across growing seasons.
4. **Mobile App or Dashboard Integration:** A lightweight, user-friendly interface could be developed for farmers or agricultural officers to input local data and receive yield estimates.
5. **Generalization to Other Crops:** The methodology could be extended to crops like cotton, rice, or sugarcane using crop-specific seasonal windows and indicators.

14 References

1. Wheat Explorer. https://ipad.fas.usda.gov/cropexplorer/cropview/commodityView.aspx?cropid=0410000 (accessed 2025-05-08).
2. Figure 4 - Relation between Wheat Crop Production and Temperature. ResearchGate. https://www.researchgate.net/figure/Relation-between-Wheat-fig3_355740318 (accessed 2025-05-08).
3. Khaki, S.; Wang, L. Crop Yield Prediction Using Deep Neural Networks. Front. Plant Sci. 2019, 10, 621. https://doi.org/10.3389/fpls.2019.00621.
4. You, J.; Li, X.; Low, M.; Lobell, D.; Ermon, S. Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data. Proc. AAAI Conf. Artif. Intell. 2017, 31 (1). https://doi.org/10.1609/aaai.v31i1.11172.
5. (PDF) Random Forest for Rice Yield Mapping and Prediction Using Sentinel-2 Data with Google Earth Engine. ResearchGate. https://doi.org/10.1016/j.asr.2022.06.073.
6. Iqbal, N.; Shahzad, M. U.; Sherif, E.-S. M.; Tariq, M. U.; Rashid, J.; Le, T.-V.; Ghani, A. Analysis of Wheat-Yield Prediction Using Machine Learning Models under Climate Change Scenarios. Sustainability 2024, 16 (16), 6976. https://doi.org/10.3390/su16166976.

7. Ali, A. M.; Abouelghar, M.; Belal, A. A.; Saleh, N.; Yones, M.; Selim, A. I.; Amin, M. E. S.; Elwesemy, A.; Kucher, D. E.; Maginan, S.; Savin, I. Crop Yield Prediction Using Multi Sensors Remote Sensing (Review Article). Egypt. J. Remote Sens. Space Sci. 2022, 25 (3), 711–716. https://doi.org/10.1016/j.ejrs.2022.04.006.