# BCSE209L – MACHINE LEARNING

## CROP YIELD PREDICTION


**22BDS0029 SORNALAKSHMI G**

**22BDS0023 PRIYADHARSHINI K**

**22BDS0155 R PRIYADARSHINI**


**SUBMITTED TO**

**AARTHY S.L**

**B.Tech.**

*in*

**Computer Science and Engineering**

**(with specialization in Data Science)**

## School of Computer Science and Engineering


September 2025

# TABLE OF CONTENTS

# 1. INTRODUCTION:

Crop yield prediction using machine learning represents a significant leap forward in agricultural technology, shifting the industry from reactive to predictive management. This approach utilizes algorithms to analyze vast and complex datasets that traditional methods cannot easily handle. By integrating diverse data sources—such as historical meteorological records, soil nutrient profiles, satellite imagery indicating plant health (NDVI), and data on irrigation and fertilizer application—machine learning models can identify critical patterns and non-linear relationships. Techniques ranging from regression models and Random Forests to more complex neural networks are trained on this historical data. The resulting models can then forecast future yields with greater accuracy, enabling farmers to optimize resource allocation, mitigate risks from climate volatility, and improve overall farm profitability, which in turn contributes to stabilizing the global food supply.

## 1.1. BACKGROUND:

The background of crop yield prediction has evolved from subjective art to data-driven science. For centuries, forecasting relied on farmers' personal experience and visual observation. This later evolved into labor-intensive statistical methods, such as physically harvesting and measuring small "crop-cutting" plots to estimate the yield of a larger area. In the late 20th century, complex biophysical simulation models (like CERES and WOFOST) were developed to mathematically mimic plant growth, but these models are extremely difficult to calibrate and require extensive local data. The true shift occurred with the advent of remote sensing—satellites like Landsat and MODIS—which provided massive, consistent datasets on vegetation health, weather, and land surface conditions. This "big data" was too complex for traditional models, paving the way for machine learning algorithms, which can effectively identify non-linear patterns and relationships within these diverse datasets to produce more accurate and scalable yield forecasts.

## 1.2. MOTIVATION:

The primary motivation for this project is to address the critical need for accuracy and efficiency in agriculture. With a growing global population, increasing climate volatility, and the rising cost of resources, traditional farming methods based on intuition are no longer sufficient. Accurate yield prediction is essential for ensuring global food security by allowing governments to anticipate and prevent food shortages. For farmers, it is an economic tool that helps them mitigate financial risk, secure better prices, and optimize planting strategies. Finally, it drives sustainability by enabling precision agriculture, allowing for the precise application of water, fertilizer, and pesticides, which reduces waste, lowers environmental impact, and improves profitability.

## 1.3. SCOPE:

This project aims to build and evaluate a machine learning model to predict crop yield for a specific crop and region. The scope includes:

**1.Data Collection:** Gathering historical data on crop yield, weather patterns, and satellite imagery (like NDVI).

**2.Model Development:** Preprocessing the data, engineering relevant features, and training several machine learning regression algorithms (e.g., Random Forest, Gradient Boosting).

**3.Evaluation:** Comparing the models using metrics like Root Mean Squared Error (RMSE) to identify the most accurate predictor.The final deliverable will be the trained model and a report analyzing its performance and identifying the most influential factors (e.g., rainfall, temperature) on yield. This project will **not** include building physical sensors or providing real-time farm management advice.

## 2. LITERATURE REVIEW

| S. NO | **Paper Title / Focus Area** | **ML Model(s) Used (or Reviewed)** | **Dataset / Key Features** | **Key Performance Metric & Result** | **Research Gap / Motivation** |
|---|---|---|---|---|---|
| **1** | **Crop Yield Prediction Using ML and DL Models** (2024) | Random Forest (**RF**), SVR, DNN, **LSTM**, ANN. | **Dataset:** Irish Maize and Potato yield data. **Features:** Weather data (temperature, rainfall). | **Metrics:** Mean Square Error (**MSE**), **$R^2$ Score**. | To **evaluate and leverage Deep Learning (LSTM/CNN)** to capture complex **temporal dependencies** in time-series weather data for improved forecasting over traditional ML. |

| S. NO | **Paper Title / Focus Area** | **ML Model(s) Used (or Reviewed)** | **Dataset / Key Features** | **Key Performance Metric & Result** | **Research Gap / Motivation** |
|---|---|---|---|---|---|
| **2** | **A Machine Learning Model for Crop Yield Prediction Using Remote Sensing Data** (2025) | Random Forest, **SVM**, Decision Trees, Linear Regression. | **Dataset:** Multi-year **Satellite Imagery** integrated with **Precipitation** data. **Features:** Vegetation Indices, Rainfall. | **Metrics:** MAE and **RMSE**. **Key Finding:** The **SVM** model showed superior performance. | To **overcome the limitations of costly, time-consuming ground-based data collection** by integrating high-dimensional, large-scale **Remote Sensing (RS)** data for scalable prediction. |
| **3** | **Integrated Approach Combining Multi-Source Remote Sensing Data** (Recent) | K-Nearest Neighbors (**KNN**), Random Forest (**RF**), Gradient Boosting. | **Crop:** Rice in the Cauvery Delta Region. **Features:** Multi-source **Sentinel-1 (SAR)** and **Sentinel-2 (Optical)** data. | **Best Performance: KNN** ($\text{R}^2 = 0.87$, $\text{RMSE} = 318 \text{ kg/ha}$). | To **improve prediction detail and robustness** by addressing the single-source limitation through **fusing diverse remote** |

| S. NO | **Paper Title / Focus Area** | **ML Model(s) Used (or Reviewed)** | **Dataset / Key Features** | **Key Performance Metric & Result** | **Research Gap / Motivation** |
|---|---|---|---|---|---|
| | | | | | **sensing data modalities** (Active SAR and Passive Optical). |
| **4** | **STATISTICAL AND MACHINE LEARNING METHODS...** (2020) | **Random Forests (RF)**, ANN, Decision Trees (DT), MLR. | **Crop:** Seventeen cash crops in Southwestern Ontario. **Features:** High-resolution **Soil properties** and **topographic characteristics**. | **Best Performance: Random Forests** achieved an **R-squared value of 0.93**. | To **establish the most effective model for precision agriculture at the field level** by quantifying the relationship between high-resolution **soil/topographic characteristics** and yield variability. |
| **5** | **A Comprehensive Review... With Special Emphasis** | ANN, Random Forest (RF), SVR, Deep Learning | **Focus:** Palm Oil. **Features:** Average Weight of Fruit Bunches (**ABW**), | **Metrics Reviewed:** MAE, MAPE, MSE, and CorrCoef. | To **provide a systematic review and identify current** |

| S. NO | **Paper Title / Focus Area** | **ML Model(s) Used (or Reviewed)** | **Dataset / Key Features** | **Key Performance Metric & Result** | **Research Gap / Motivation** |
|---|---|---|---|---|---|
| | **on Palm Oil** (2021) | (DL), CNN. (Review) | **BUNCHHA**, Remote Sensing. | | **gaps** in ML application specifically for an **economically crucial and complex perennial crop (Palm Oil)**, which has unique prediction challenges. |
| **6** | **CROPS YIELD PREDICTION...** (West African Countries) (2023) | Three distinct machine learning models. | **Crop:** Six crop yields, including **rice**. **Features:** Statistical data of environmental/agronomic elements in West Africa. | **Performance:** Mentions specific yield prediction results, e.g., **0.162 kg/ha**. | To **address food scarcity and data-driven farming challenges in under-represented, climate-vulnerable regions** (West Africa) using available statistical data. |

| S. NO | **Paper Title / Focus Area** | **ML Model(s) Used (or Reviewed)** | **Dataset / Key Features** | **Key Performance Metric & Result** | **Research Gap / Motivation** |
|---|---|---|---|---|---|
| **7** | **Machine Learning-Based Crop Yield Prediction in South India** (2024) | **Extra Trees Regressor**, Linear, Neighbors-based models. | **Crop:** Rice, Sorghum, Cotton, Sugarcane, and Rabi crops in South India. **Features:** Weather, Soil, and Crop historical data. | **Best Performance:** Extra Trees Regressor ($\text{R}^2 = 0.9615$). | To **develop a high-accuracy, regional-specific model** for multiple major cash crops in a **climatically diverse and agriculturally intensive area (South India)**. |
| **8** | **Advances in Machine Learning...** (2025) | Random Forests (RF), Gradient Boosting Decision Trees (**GBDT**), ANNs. (Review) | **Features:** Soil metrics, climatic variables, and general crop characteristics. | **Key Finding:** A cited ensemble learning model achieved $\text{R}^2 = 0.96$ and $\text{MSE} = 42,963$. | To **systematically review recent advancements in ensemble and deep learning methods** and **identify key research limitations** (e.g., data |

| S. NO | **Paper Title / Focus Area** | **ML Model(s) Used (or Reviewed)** | **Dataset / Key Features** | **Key Performance Metric & Result** | **Research Gap / Motivation** |
|---|---|---|---|---|---|
| | | | | | heterogeneity, generalizability) for future direction. |
| **9** | **Comprehensive Analysis of Crop Yield Prediction Using Deep Learning and Remote Sensing...** (2025) | **CNNs** and **LSTMs**, Hybrid models, Attention mechanisms. (Proposed Framework/ Review) | **Features:** Satellite imagery (MODIS, Sentinel), **Vegetation indices**, Land surface temperature, Soil moisture. | **Goal:** Proposes a new DL framework to enhance prediction accuracy and applicability. | To **overcome the limitations of traditional ML in handling complex spatiotemporal data** by proposing and evaluating **advanced Deep Learning (DL) architectures and hybrid frameworks**. |
| **10** | **Crop Yield Estimation using Machine** | Linear Regression (**LR**), Decision Tree (**DT**), Random | **Crop:** Cocoa, Oil Palm, Rice, Rubber. **Features:** Environmental parameters like | **Goal:** Paper outlines a methodology to improve | To **evaluate the optimal predictive power of** |

| S. NO | **Paper Title / Focus Area** | **ML Model(s) Used (or Reviewed)** | **Dataset / Key Features** | **Key Performance Metric & Result** | **Research Gap / Motivation** |
|---|---|---|---|---|---|
| | **Learning** (2025) | Forest (**RF**), Gradient Boosting (**GB**). | Precipitation, Specific Humidity, Temperature. | model performance with feature engineering. | **various fundamental environmental factors** for diverse cash crops and **optimize feature selection** for simpler, more robust models. |

## 3. OBJECTIVES

**Objective 1:**

**To Develop a Predictive Model:** To design, build, and train a robust machine learning regression model (e.g., Random Forest, Gradient Boosting, or LSTM) capable of accurately forecasting crop yield by integrating diverse data sources, including historical weather patterns, soil characteristics, and satellite remote sensing data (like NDVI).

**Objective 2:**

**To Evaluate and Compare Model Performance:** To systematically evaluate and compare the performance of multiple machine learning algorithms using standard regression metrics, such as Root Mean Squared Error (RMSE) and Coefficient of Determination ($R^2$). The goal is to identify and validate the most accurate and reliable model for the specific crop and region being studied

**Objective 3:**

**To Identify Key Influential Factors:** To analyze the best-performing model to identify and quantify the most significant environmental and agronomic factors that influence crop yield.

This involves using feature importance techniques to understand which variables (e.g., rainfall during the flowering stage, maximum temperature, soil nitrogen) are the strongest predictors, thereby providing actionable insights for farmers and policymakers.

## 4. HARDWARE SPECIFICATIONS

The required hardware is based on the project scale (single farm, and for the nation) and model complexity. A typical configuration for a moderate-sized project could include:

CPU: A modern multi-core CPU is appropriate. An Intel Core i7 or i9, or an AMD Ryzen 7 or 9, with at least 8 cores is important for efficient data preprocessing, and results in high-level computing of multiple tasks.

GPU: This is the most critical piece of hardware when training deep learning models. The industry standard is an NVIDIA GPU with CUDA support.

Entry level: NVIDIA GeForce RTX 3060 (12 GB VRAM)

Mid-level/professional level: NVIDIA GeForce RTX 4070 or RTX 3080 (12-16 GB VRAM)

Top of the line/cloud level: NVIDIA A100 or H100 (40-80 GB VRAM) for large datasets and complex models.

RAM: At least 32 GB of RAM is a recommendation to process large datasets without slowing down, 64 GB ram or larger is needed for very large datasets (terabytes).

Storage: A very fast SSD, preferably a NVMe SSD with at least 1 TB, for quick access and loading of data. A separate HDD will be used for older long term access data.

## 5. SOFTWARE SPECIFICATIONS

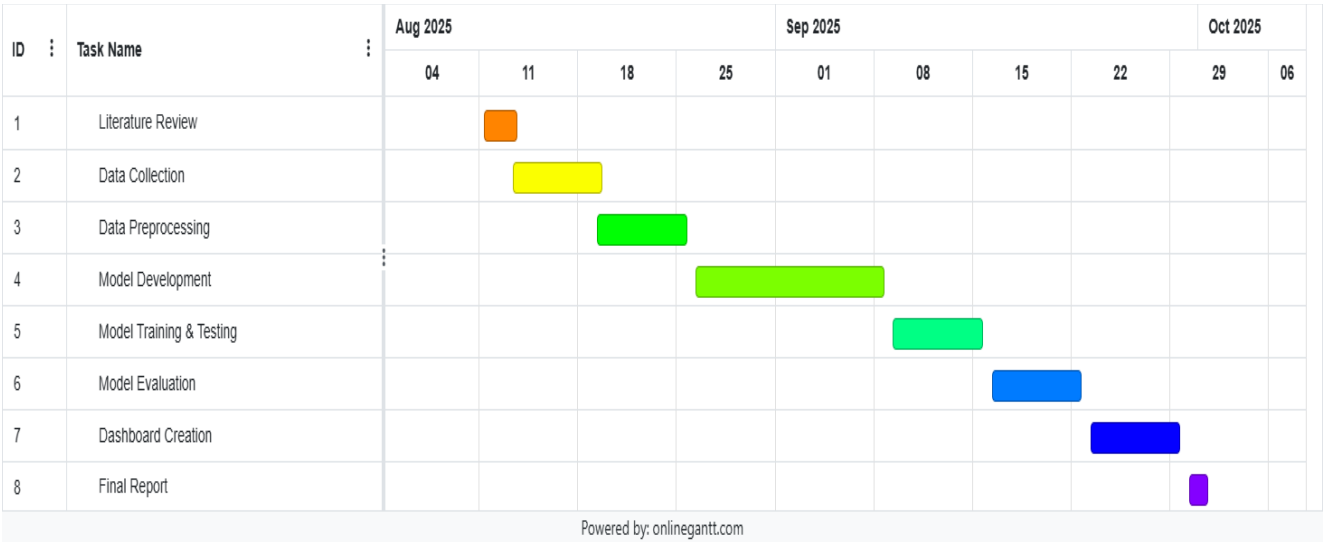
The software stack will provide the environment for the development, training, and deployment of the machine learning model.

Operating System: A Linux-based operating system like Ubuntu 20.04/22.04 LTS is the preferred operating system due to the stability, security, and ease of integration with machine learning packages. Windows (with WSL2) or Mac can also be used.
Programming Language: Python (version 3.8+) is the predominant programming language for machine learning, as it is simple to learn and supports a large number of libraries.
**Core ML/DL Frameworks.**

TensorFlow or PyTorch: These two frameworks are the premier deep learning frameworks for developing and training neural networks (e.g., LSTMs, CNNs) which are frequently used to analyze satellite imagery and time-series weather data.
ConScikit-learn: This will be necessary for traditional machine learning models (e.g., Random Forest, Gradient Boosting) and for some data preprocessing, evaluating models, and hyperparameter tuning.
**Data Manipulation & Analysis Libraries**.
Pandas: For loading, cleaning, and manipulating structured data (e.g., weather records, soil data).
NumPy: High-performance numerical computation.
**Data Visualization Libraries.**
Matplotlib & Seaborn: For visualizing charts and graphs to visualize the distributions of the data and performance of the model.
Jupyter Notebook/Jupyter Lab: A perfect environment to explore data, prototype models, and visualize much of the previous.
VS Code or PyCharm: Integrated Development Environments (IDEs) for writing and maintaining larger code bases.

## 6. GANTT CHART

The project is structured into eight distinct phases, managed chronologically to ensure a systematic progression from research to final reporting. The entire project is scheduled to be completed over approximately two months, starting in early August 2025 and concluding in early October 2025.




## 7. WORK BREAKDOWN STRUCTURE

The WBS developed for this project is divided into eight primary phases: Literature Review, Data Collection, Data Preprocessing, Model Development, Model Training & Testing, Model Evaluation, Dashboard Creation, and Documentation & Final Report. Each of these phases is further broken down into specific subtasks.This detailed structure serves as the basis for the project's schedule and management


# 8. REQUIREMENT ANALYSIS

## 8.1. FUNCTIONAL REQUIREMENTS

Data Collection: The system will automatically collect and analyze data from a variety of sources that includes:

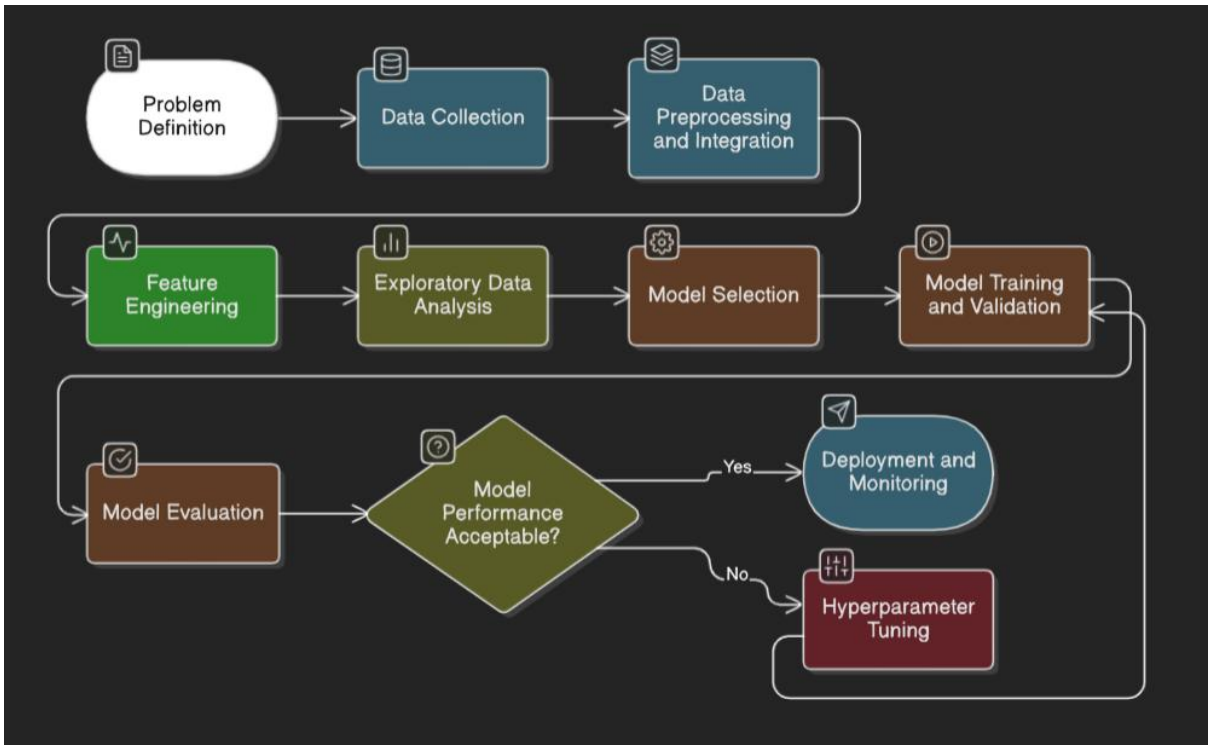
**Weather API's:**For real-time and historical weather information(temperature, rainfall, humidity).
**Satellite Imaging Vendors**: For geospatial data such as NDVI (Normalized Difference Vegetation Index).
**IoT Sensors**: For soil moisture, pH and nutrient levels on the field.Manual Upload: Enable users to upload historical yield data and farm record in a CSV or Excel format.
**Data Preprocessing**: In an automated manner, the system will identify clean, normalize and convert the incoming data into machine learning ready format. This will include addressing missing values and aligning time-series data.
**Model Training & Retraining**: The system will be able to train machine learning models with the cleaned dataset. scheduled retraining should also facilitate additional incorporation of data while the model remains accurate.
**Prediction Generation**: The main function of the system will be to generate predictions of Y, crop yield (bushels), for a specific location and time duration. The output will be a numerical value (e.g. tons per ha.) with an associated confidence interval.

### 8.2. NON FUNCTIONAL REQUIREMENTS

These categorize the performance features and operating characteristics of the system.

**Accuracy:** The predictive model must attain at least a minimum predetermined accuracy descriptor (e.g., show a Root Mean Square Error (RMSE) below a certain number or R 2 field above 0.85, etc.)
**Performance**: The system needs to make a prediction for a specific farm in 10 seconds from the time a user requests a prediction. The dashboard must load under 3 seconds.
**Scalability**: The system architecture must be able to support a 50% increase in the number of users and number of data sources over a period of two years without a meaningful degradation of performance.
**Reliability & Availability:** The system must meet 99.5% uptime. It must be therefore available during important planting and harvesting periods.
Security: Any user data (especially farm locations and proprietary data) must be secured, whether at rest or in motion, through encryption. The system must implement role-based access.

## 9. WORKFLOW MODEL




## Step 1: Problem Definition and Scoping

Before any data is collected, you must define the precise goal.

**Target Variable:** What exactly are you- predicting? (e.g., Rice yield in tons/hectare).

**Spatial Resolution:** What is the geographic unit of prediction? (e.g., farm-level, district-level, or state-level). Farm-level is more complex but more actionable.

**Temporal Resolution:** When is the prediction made? (e.g., end-of-season forecast, or an intra-season forecast made mid-growing season).

**Specific Crop:** The model will be specific to one crop (e.g., maize, wheat, or soybean), as different- crops have different environmental responses.

## Step 2: Data Collection

This is often the most challenging step, as data must be gathered from diverse sources. No single dataset typically contains all necessary information (Engen et al., 2021).

**Historical Yield Data (Target):** This is the ground-truth data you are training the model to predict. It can be sourced from government agricultural agencies, farm records, or organizations like the FAO (Food and Agriculture Organization).

**Weather (Meteorological) Data:** This is a primary driver of yield. Key features include daily temperature (min, max, average), precipitation, solar radiation, and humidity (Kuradusenge et al., 2023).

**Soil Data:** This includes soil properties for the specific locations. Important features are soil type, pH, organic carbon content, and nutrient levels (Nitrogen, Phosphorus, Potassium - NPK).

**Remote Sensing (Satellite) Data:** This is crucial for monitoring plant health in real-time. Data from satellites like Sentinel-2 or Landsat is used to calculate Vegetation Indices (VIs).

**Agronomic/Management Data:** (If available) This includes data on farming practices, such as fertilizer application rates, irrigation type and frequency, and planting dates.

## Step 3: Data Preprocessing and Integration

Collected data is raw, messy, and disparate.

**Handling Missing Data:** Weather stations have downtime, and satellite images can be obscured by clouds. Use techniques like temporal interpolation (for weather) or spatial interpolation to fill gaps.

**Data Integration:** This is a critical challenge. All data must be aggregated to the same spatio-temporal unit. For example, for a *district-level* prediction for the *2024 season*, you must:Get the single, final yield value for that district in 2024.Average all daily weather data across all stations within that district for the entire growing season.Average all soil sample data within the district.Average all satellite pixel data (e.g., NDVI) within the district's boundaries over the growing season.

**Data Cleaning:** Remove outliers (e.g., a yield data entry that is 100x the average).

## Step 4: Feature Engineering

Raw data is often not as predictive as engineered features.

**Weather Features:** Instead of just "average annual temperature," create features relevant to crop growth stages, such as Total precipitation during the vegetative stage.Average temperature during the flowering/reproductive stage."Stress Days": Count of days where the temperature exceeded a critical threshold (e.g., > 35°C).

**Remote Sensing Features:** The most important feature is the **Normalized Difference Vegetation Index (NDVI)**. NDVI is calculated from satellite imagery (using red and near-infrared bands) and is a strong proxy for plant health, density, and photosynthetic activity (Wang et al., 2024). You can engineer features like "peak NDVI," "average NDVI during mid-season," or the "time to peak NDVI."

## Step 5: Exploratory Data Analysis (EDA)

Analyze the processed data to find patterns.

**Correlation Analysis:** Plot a heatmap to see which features (e.g., rainfall, NDVI) have the strongest positive or negative correlation with the target (yield).

**Feature Importance:** A preliminary Random Forest model can be trained to identify the most influential variables. For example, studies have shown that factors like nitrogen fertilizer application and climate variables are often top predictors (Jeong et al., 2016).

**Visualization:** Create maps showing yield variations by region and plots showing yield trends over time.

## Step 6: Model Selection

Start with a simple model and move to more complex ones.

**Baseline Model:** A simple Multiple Linear Regression (using key features like rainfall and temperature) to establish a performance baseline.

**Tree-Based Models:** These are very popular and highly effective for this type of tabular data.

**Random Forest (RF):** A robust model that handles non-linear relationships well and is less prone to overfitting. It is frequently cited as a top-performing model for yield prediction (Jeong et al., 2016; Kuradusenge et al., 2023).

**Gradient Boosting (XGBoost, LightGBM):** Often provides the highest accuracy on tabular data by iteratively correcting the errors of previous models

**Deep Learning Models:**

**LSTMs (Long Short-Term Memory):** A type of Recurrent Neural Network (RNN) that is excellent if you use the full time-series of weather/NDVI data without aggregation, as it can learn temporal patterns (Archana & Kumar, 2023).

**CNN (Convolutional Neural Network):** Can be used to extract features directly from satellite images.

## Step 7: Model Training and Validation

How you split your data is critical.

**Train/Test Split:** Do not split data randomly. Crop yield data is *temporal* (sequential by year) and *spatial*.

**Validation Strategy:** A *temporal* split is essential. Train your model on data from, for example, 2000-2018, and then test its performance on "unseen" future years (e.g., 2019-2020). This mimics a real-world forecasting scenario. You can also use *k-fold cross-validation* for a more robust performance estimate.

## Step 8: Model Evaluation

Since this is a regression task, you will use regression metrics to evaluate the model's accuracy.

**$R^2$ (R-squared):** The coefficient of determination. It measures the proportion of the variance in the yield that is predictable from the features. An $R^2$ of 0.85 means the model can explain 85% of the variability in crop yield.

**Root Mean Squared Error (RMSE):** This is the most common metric. It gives you the average error of the model in the same units as the target variable (e.g., an RMSE of 0.5 means the model is, on average, off by 0.5 tons/hectare).

**Mean Absolute Error (MAE):** Similar to RMSE, but it is less sensitive to large outlier errors.

## Step 9: Hyperparameter Tuning

Once you have selected your best-performing model (e.g., Random Forest), you can use techniques like Grid Search or Randomized Search to fine-tune its internal settings (e.g., the number of trees, the depth of each tree) to squeeze out the best possible performance.

## Step 10: Deployment and Monitoring

The final step is to put the model into production.

**Deployment:** The model can be deployed as a web application or an API. A farmer or policymaker could input their region and receive a yield forecast based on the current season's data.

**Monitoring:** The model's predictions must be continuously monitored against actual yields as they are reported. Models can become "stale" as climate patterns shift or new farming technologies are introduced (a concept called *model drift*). The model should be retrained with new data every 1-2 years.

## 10. MODULE DESIGN

## 11. REFERENCES:


**1.Comprehensive Review (Broad Survey)** Bou-Hamad, I., & El-Khoury, H. (2024). Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. *Frontiers in Plant Science*, *15*. This is an excellent starting point, as it reviews a large body of work and discusses the most common machine learning (ML) and deep learning (DL) models, data sources, and challenges.

**2.Systematic Review of ML/DL (Recent Trends)** Rahman, M. A., Shoaib, M., & Ashraf, M. (2024). Crop Yield Prediction Using Machine Learning: An Extensive and Systematic Literature Review. *Sustainability*, *16*(10), 4005. This paper systematically reviews 184 papers to identify the most-used models (like Random Forest, Gradient Boosting, CNN, and LSTM) and evaluation metrics (RMSE, $R^2$), providing a clear snapshot of the current research landscape.

**3.Focus on Remote Sensing & UAVs (Data Source)** Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N. H., & Islam, N. (2022). A Systematic Literature Review on Crop Yield Prediction with Deep Learning and Remote Sensing. *Remote Sensing*, *14*(9), 1990. This review focuses specifically on how deep learning models use remote sensing data, particularly from satellites and Unmanned Aerial Vehicles (UAVs), for yield prediction.

**4.Foundational Deep Learning Paper** Khaki, S., & Wang, L. (2019). Crop Yield Prediction Using Deep Neural Networks. *Frontiers in Plant Science*, *10*, 621. This is a highly-cited paper that demonstrates the superior performance of deep neural networks (DNNs) compared to other methods like LASSO and shallow neural networks, setting a benchmark for deep learning approaches.

**5.Focus on Random Forest (Common Model)** Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., ... & Kim, S. H. (2016). Random Forests for Global and Regional Crop Yield Predictions. *PLOS ONE*, *11*(6), e0156571. A foundational paper demonstrating the effectiveness and versatility of the Random Forest algorithm for predicting crop yields at both regional and global scales.

**6.Combining Data Sources (Practical Implementation)** Shah, P., Sahu, S., & Subudhi, B. N. (2023). Improving Wheat Yield Prediction with Multi-Source Remote Sensing Data and Machine Learning in Arid Regions. *Remote Sensing*, *17*(5), 774. This paper details a practical approach of combining multi-source data (like remote sensing indices and climate data) with models like Random Forest (RF) and Gradient Boosting (GB) to predict wheat yield.

**7.Focus on Explainability & Hybrid Models** G. S, S., & S. A, V. (2023). Crop Yield Prediction Using Improved Random Forest: A Hybrid Approach. *ITM Web of Conferences*, *56*, 02007. This article presents a hybrid model using an "Improved Random Forest" and highlights the importance of feature selection, which is a key part of building an interpretable and accurate model.

**8.Time-Series Focus with LSTMs** Sun, J., Wu, M., Di, L., & Sun, Z. (2020). Integrating Remote Sensing and Weather Data for Hybrid Rice Yield Prediction Using a Long Short-Term Memory (LSTM) Network. *Remote Sensing*, *12*(23), 3928. This study is a great example of using LSTMs, a type of recurrent neural network (RNN), to handle the time-series nature of both satellite vegetation indices (like NDVI) and weather data throughout a growing season.

**9.Focus on XGBoost (Popular Model)** Pant, J., Pant, R. P., Singh, M. K., & Singh, D. (2021). Predicting Annual Crop Yields in India's States: Leveraging XGBoost Techniques for a Web-Based Machine Learning. *International Journal of Research and Analytical Reviews (IJRAR)*, *8*(1), 724-733. This paper demonstrates the application of XGBoost, a powerful gradient-boosting algorithm, for predicting yields and discusses its robustness in handling diverse environmental factors.

**10.Integrating Deep Learning and Process-Based Models** Wang, N., Wang, E., Wang, X., Yang, N., & Zhang, Z. (2020). A deep learning-based hybrid model for crop yield prediction. *Computers and Electronics in Agriculture*, *178*, 105753. This paper explores an advanced hybrid approach, combining a process-based crop model (which simulates plant growth) with a deep learning model to improve prediction accuracy by leveraging the strengths of both methods.