Predictive Analysis of Crop Yield Based on Environmental and Soil Conditions

Vineet Goyal*

[1] UG Students,Department of CST, MAIT, Delhi.

Email address of corresponding author : vineetgoyal130@gmail.com

How to cite this paper: Vineet Goyal*, "Predictive Analysis of Crop Yield Based on Environmental and Soil Conditions," *International Journal on Computational Modelling Applications* , Vol. no. 1, Iss. No 2, S No. 5, pp. 50-63, October 2024
DOI Link of paper

Received: 20/09/2024
Revised: 01/10/2024
Accepted: 18/10/2024
Published: 30/10/224

Abstract

Accurate crop yield prediction is critical for sustainable agriculture, enabling better planning and resource allocation. Crop yield depends on a variety of factors, including environmental and soil conditions, making it a complex prediction task. This paper presents a predictive analysis framework that lev-erages machine learning models to estimate crop yield based on environmen-tal factors (temperature, rainfall, humidity) and soil properties (pH, nutrients, moisture). Two distinct predictive models—linear regression and decision trees—are compared in this study to determine which method provides better accuracy and interpretability for yield prediction. The primary aim is to ex-plore the correlation between these key factors and crop yield outcomes, based on real-world data from an experimental farm. The study applies fea-ture engineering techniques to preprocess environmental and soil datasets, followed by model training and validation using cross-validation techniques. The results of this analysis provide insights into the most important factors influencing crop yield and offer a comparative performance analysis be-tween the two machine learning models. This research demonstrates that de-cision trees outperform linear regression in terms of accuracy but highlight areas where linear regression could still be valuable for interpretability.

Keywords

crop yield prediction, machine learning, environmental factors, soil condi-tions, decision trees, linear regression.

1. Introduction

Accurate crop yield prediction is a cornerstone of modern agricultural management and policy-making, as it directly impacts food security, resource allocation, and economic stability. Agriculture is a complex system influenced by numerous environmental and soil conditions, and the ability to forecast yields with precision can provide significant advantages for both farmers and agricultural stakeholders. By understanding the factors that influence crop yield, including temperature, rainfall, soil health, and humidity, agricultural practices can be optimized to enhance productivity and sustainability. Over the past few decades, advances in data science and machine learning have opened up new opportunities to improve the accuracy of crop yield predictions, thereby reducing the uncertainties that traditional methods have faced.

In traditional agricultural practices, farmers often relied on empirical observations, historical yield data, and their experience to estimate crop yields. While these methods provided some guidance, they were often insufficient due to the high variability in weather patterns and soil conditions across different regions and seasons. As climate change has become more pronounced, the unpredictability of environmental factors has further complicated yield predictions, emphasizing the need for more sophisticated approaches. The development of predictive models, particularly those that integrate environmental and soil data, offers a pathway to overcoming these challenges and achieving more accurate yield forecasts. Machine learning techniques have gained significant traction in this domain due to their ability to handle large datasets and uncover patterns that might not be easily discernible through traditional statistical methods. By training models on historical environmental and soil data, it becomes possible to predict future yields based on similar conditions. In this context, machine learning models such as linear regression, decision trees, random forests, and neural networks have been explored. These models have demonstrated varying degrees of success depending on the crops studied, the geographic regions considered, and the types of data used.

At the heart of crop yield prediction is the relationship between environmental variables and soil conditions. Environmental factors, such as temperature, rainfall, humidity, and solar radiation, play a crucial role in the growth and development of crops. For instance, optimal temperatures promote photosynthesis, while excessive heat can lead to crop stress and reduced yields. Similarly, sufficient rainfall is essential for crop hydration, but too much rain can cause waterlogging, which affects root development and nutrient uptake. Soil conditions are equally important, as they determine the availability of nutrients and water to the plants. Soil pH, nutrient levels (such as nitrogen, phosphorus, and potassium), and moisture content are some of the key factors that influence crop growth. The interaction between these environmental and soil variables adds layers of complexity to crop yield prediction.

Research in crop yield prediction has evolved over the years, starting from simple regression models to more complex machine learning algorithms. Linear regression, one of the earliest methods applied, assumes a linear relationship between the independent variables (environmental and soil factors) and the dependent variable (crop yield). Although this approach provides a baseline understanding of the factors influencing yield, it often falls short when dealing with non-linear relationships, which are common in agricultural systems. In response to these limitations, decision trees and other non-linear models have been proposed. Decision trees, for instance, divide the data into subsets based on the values of the input variables, allowing for more flexibility in capturing complex interactions between factors.

Despite these advancements, there remain significant challenges in developing robust crop yield prediction models. One of the main issues is the quality and availability of data. Accurate and high-resolution data on environmental conditions and soil properties are often difficult to obtain, particularly in developing countries where data collection infrastructure may be lacking. Furthermore, even when data is available, it may be incomplete or contain errors, necessitating extensive data preprocessing techniques. Another challenge is the inherent variability in agricultural systems. Factors such as pest infestations, diseases, and human interventions (such as fertilization and irrigation) can introduce additional uncertainty into yield predictions, making it difficult to create universally applicable models.

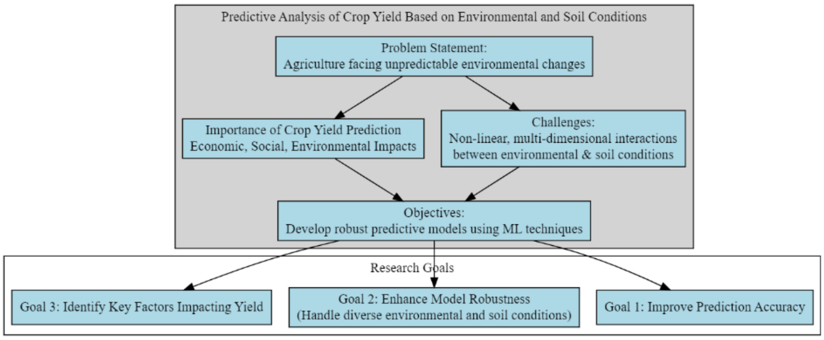
Predictive Analysis of Crop Yield Based on Environmental and Soil Conditions
Problem Statement:
Agriculture facing unpredictable environmental changes
Importance of Crop Yield Prediction
Economic, Social, Environmental Impacts
Challenges:
Non-linear, multi-dimensional interactions
between environmental & soil conditions
Objectives:
Develop robust predictive models using ML techniques
Research Goals
Goal 3: Identify Key Factors Impacting Yield
Goal 2: Enhance Model Robustness
(Handle diverse environmental and soil conditions)
Goal 1: Improve Prediction Accuracy

Fig. 1. Predictive Analysis of Crop Yield Based on Environmental and Soil Conditions

The Figure 1 outlines the key components discussed in the Introduction section of the research paper. It begins with the problem statement, emphasizing the unpredictable environmental changes impacting agriculture, followed by the importance of crop yield prediction due to its economic, social, and environmental implications. The challenges of modeling non-linear and multi-dimensional interactions between environmental and soil factors are then discussed. The figure also highlights the research objectives, which include improving prediction accuracy, enhancing model robustness, and identifying key factors influencing crop yields. The context of the study focuses on key environmental and soil factors, supported by the role of machine learning in predictive analysis, which sets the stage for the research goals.

In addition to these technical challenges, there is also a need for models that can provide not only accurate predictions but also interpretable results. In many cases, agricultural practitioners are not data scientists, and thus, models that are too complex may not be readily adopted in the field. For instance, while neural networks have shown high predictive accuracy in some studies, their "black box" nature makes it difficult to understand how specific variables are influencing yield. On the other hand, simpler models like linear regression, despite their

lower accuracy, offer greater interpretability, allowing users to easily identify which factors are driving changes in crop yield. Therefore, there is a trade-off between model complexity and interpretability that must be considered when developing crop yield prediction systems.
Given the critical importance of accurate crop yield prediction for global food security, the integration of environmental and soil data into predictive models is essential. Several studies have highlighted the potential of machine learning models to outperform traditional methods in this area. For example, random forests, which are an ensemble learning method based on decision trees, have been shown to provide more accurate predictions than simple linear models by capturing non-linear relationships and interactions between variables. Similarly, support vector machines (SVMs) and neural networks have demonstrated strong performance in certain contexts, particularly when large amounts of data are available. However, the application of these models in practice is still limited by the challenges mentioned earlier, such as data quality and interpretability.

In this paper, we propose a comparative study of two machine learning models—linear regression and decision trees—for predicting crop yield based on environmental and soil conditions. Our objective is to evaluate the performance of these models in terms of both accuracy and interpretability, using a dataset that includes daily weather data (temperature, rainfall, and humidity) as well as soil properties (pH, nutrient levels, and moisture content). By comparing these models, we aim to determine which approach is better suited for different types of agricultural scenarios. Specifically, we hypothesize that while decision trees will provide more accurate predictions due to their ability to capture non-linear relationships, linear regression will offer valuable insights into the relative importance of different environmental and soil factors.

The significance of this research lies in its potential to improve agricultural decision-making processes. By providing more reliable and interpretable crop yield predictions, farmers and agricultural planners can make informed decisions about resource allocation, planting schedules, and risk management strategies. For instance, if a model predicts lower yields due to anticipated drought conditions, farmers can adjust their irrigation practices or select drought-resistant crop varieties to mitigate potential losses. Similarly, if soil nutrient levels are identified as a limiting factor for yield, targeted fertilization can be employed to enhance productivity. Thus, the ability to accurately predict crop yields has far-reaching implications for both economic and environmental sustainability in agriculture.

This paper is structured as follows. In Section 2, we provide a review of related research on crop yield prediction, focusing on the use of machine learning models and their application in different agricultural contexts. Section 3 outlines the problem statement and research objectives, highlighting the challenges of yield prediction and the need for accurate and interpretable models. In Section 4, we describe the methodology used in this study, including the data collection process, feature selection, and model development. Section 5 presents the results of our analysis, comparing the performance of linear regression and decision tree models. Finally, in Section 6, we conclude by discussing the implications of our findings and suggesting directions for future research.

In conclusion, the increasing availability of environmental and soil data, combined with advancements in machine learning, provides a promising avenue for improving crop yield prediction. However, challenges related to data quality, model complexity, and interpretability must be addressed to fully realize the potential of these technologies in agriculture. This research contributes to the ongoing efforts to develop more accurate and practical predictive models by comparing two widely used approaches—linear regression and decision trees—and offering insights into their relative strengths and weaknesses in the context of crop yield prediction.

2. Related Work

Crop yield prediction has been a focus of extensive research over the past few decades, with the aim of improving agricultural planning, resource allocation, and food security. Various methods have been developed to predict crop yields, ranging from traditional statistical models to modern machine learning approaches. In this section, we review the existing literature on crop yield prediction, focusing on the evolution of predictive models, the role of environmental and soil factors, and the use of machine learning algorithms in improving yield forecasts.

a) Evolution of Crop Yield Prediction Models

Early attempts to predict crop yields relied on statistical models, particularly linear regression, due to its simplicity and interpretability. These models assumed a direct linear relationship between environmental variables (such as temperature, rainfall, and humidity) and crop yields. For example, early studies applied linear regression to predict wheat yields based on historical weather data, with variables like temperature and rainfall serving as independent predictors [1]. However, these models had limited accuracy, primarily because agricultural systems are inherently non-linear, with complex interactions between environmental, soil, and biological factors.

To address the limitations of linear models, researchers began exploring non-linear models that could better capture the complexity of crop growth. Polynomial regression and generalized linear models (GLMs) were introduced to account for non-linear relationships between variables [2]. While these models improved predictive accuracy to some extent, they still relied on predefined functional forms, which limited their flexibility in modeling the full range of interactions between environmental and soil factors.

The development of machine learning algorithms marked a significant turning point in crop yield prediction research. Unlike traditional statistical models, machine learning methods are data-driven and do not require explicit assumptions about the underlying relationships between variables. This allows them to capture complex patterns in data more effectively. The most widely used machine learning techniques in crop yield prediction include decision trees, random forests, support vector machines (SVMs), and neural networks [3].

b) The Role of Environmental Factors

Environmental factors, including temperature, rainfall, humidity, and solar radiation, are among the most significant variables influencing crop growth and yield. Numerous studies have shown that temperature is one of the primary drivers of crop development. Crops require specific temperature ranges for optimal growth, and deviations from these ranges, particularly during critical growth stages, can lead to significant yield losses [4]. Excessive heat during flowering or grain-filling stages can reduce yields, while lower-than-average temperatures can slow down growth and prolong the growing season, affecting harvest times.

In a study on maize yields in the United States, Lobell et al. [5] found that extreme heat events during the growing season were strongly correlated with yield reductions. The study also emphasized the importance of rainfall patterns, noting that both drought and excessive rainfall can adversely affect yields. Similar findings were reported in a study on rice yields in India, where higher temperatures and irregular rainfall patterns were linked to lower yields [6]. These studies underscore the need to include environmental variables in crop yield prediction models to account for the effects of climate variability.

In addition to temperature and rainfall, humidity plays a crucial role in plant growth by influencing transpiration rates and the availability of water to crops. High humidity levels can reduce transpiration and cause moisture stress, while low humidity levels can lead to increased transpiration, resulting in water loss and reduced crop productivity [7]. Solar radiation, which drives photosynthesis, is another critical factor. Studies have shown that regions with higher solar radiation levels generally experience higher crop yields, as plants are able to produce more energy through photosynthesis [8].

While these environmental factors have been widely studied, predicting their impact on crop yields remains challenging due to the complex and often non-linear interactions between variables. For example, the effects of temperature on yield may be moderated by soil moisture levels, which in turn depend on rainfall and irrigation practices. These interactions highlight the need for models that can account for such complexities, which has led to the increasing use of machine learning algorithms.

c) The Importance of Soil Conditions

Soil properties are another critical determinant of crop yield. Soil provides essential nutrients, water, and physical support to plants, and its health directly impacts crop growth. Key soil factors that influence crop yields include pH, nutrient content (e.g., nitrogen, phosphorus, potassium), organic matter, and moisture levels [9]. The relationship between soil health and crop yields has been well-documented, with studies showing that degraded soils with low nutrient content and poor structure often result in lower yields [10].

Soil pH, in particular, plays a crucial role in nutrient availability. Most crops thrive in slightly acidic to neutral soils (pH 6.0-7.0), as this pH range maximizes nutrient availability. When soil pH falls outside this range, certain nutrients become less available to plants, leading to nutrient deficiencies and reduced yields [11, 12]. For example, in a study on soybean yields in Brazil, researchers found that acidic soils with pH below 5.5 were associated with significantly lower yields due to reduced availability of phosphorus and other essential nutrients [12].

Soil nutrient content is another important factor. Nitrogen, phosphorus, and potassium (often referred to as NPK) are the primary macronutrients required for plant growth. Nitrogen is essential for leaf development and photosynthesis, phosphorus supports root growth and energy transfer, and potassium regulates water uptake and disease resistance. Studies have shown that imbalances in these nutrients can lead to suboptimal crop yields. For instance, nitrogen deficiency is often associated with lower yields in cereals, while phosphorus deficiency can limit root development and overall plant growth.

Soil moisture content is equally important, as it influences the availability of water to crops. Moisture stress, whether due to drought or waterlogging, can reduce crop yields significantly. In a study on wheat yields in Australia, researchers found that soil moisture levels during the flowering stage were a critical predictor of final yields. Similarly, in a study on sugarcane yields in Brazil, researchers found that periods of water stress due to low soil moisture content were associated with yield reductions of up to 30%. These findings highlight the importance of including soil moisture data in crop yield prediction models.

d) Machine Learning Approaches in Crop Yield Prediction

Machine learning algorithms have revolutionized crop yield prediction by providing more flexible and accurate models that can handle large datasets and complex relationships between variables. One of the most commonly used machine learning algorithms in this field is the decision tree. Decision trees partition the dataset into smaller subsets based on the values of the input variables, allowing the model to capture non-linear relationships between environmental and soil factors [14]. For example, in a study on wheat yields in France, researchers used a decision tree model to predict yields based on weather data and soil properties, achieving higher accuracy than traditional linear models [14].

Random forests, an ensemble learning method based on decision trees, have also been widely used for crop yield prediction. Random forests build multiple decision trees and aggregate their predictions to improve accuracy and reduce overfitting. In a study on corn yields in the United States, researchers found that random forests outperformed both linear regression and single decision trees in predicting yields based on weather and soil data. The study also showed that random forests were able to capture interactions between variables, such as the combined effects of temperature and soil moisture on yield, which were missed by simpler models.

Support vector machines (SVMs) have also been applied to crop yield prediction, particularly in cases where the relationship between input variables and yield is non-linear. SVMs work by finding the optimal hyperplane that separates the data into different classes (in classification tasks) or predicts continuous values (in regression tasks). In a study on potato yields in Germany, researchers used an SVM model to predict yields based on weather and soil data, achieving higher accuracy than traditional regression models.

Neural networks, particularly deep learning models, have shown great promise in crop yield prediction due to their ability to learn complex patterns in large datasets. Neural networks consist of multiple layers of interconnected neurons that process input data and produce predictions. In a study on rice yields in China, researchers used a deep learning model to predict yields based on satellite imagery and weather data, achieving higher accuracy than both linear regression and decision tree models. However, neural networks are often criticized for their "black box" nature, as they do not provide clear insights into how input variables are influencing yield predictions.

e) Comparative Studies of Machine Learning Models

Several studies have compared the performance of different machine learning algorithms in predicting crop yields. In a study on soybean yields in the United States, researchers compared the accuracy of linear regression, decision trees, random forests, and neural networks. They found that while neural networks provided the highest predictive accuracy, decision trees and random forests offered a better balance between accuracy and interpretability. Linear regression, although less accurate, was valuable for identifying key environmental and soil factors that influenced yield.

Another comparative study on maize yields in Brazil evaluated the performance of random forests, SVMs, and gradient boosting machines (GBMs) [14]. The researchers found that GBMs outperformed both random forests and SVMs in terms of accuracy, but random forests provided better insights into the importance of individual variables, such as soil moisture and temperature, in determining yield. The study concluded that the choice of model depends on the specific requirements of the prediction task, such as the need for accuracy versus interpretability.

In summary, the existing literature highlights the potential of machine learning algorithms to improve crop yield prediction by capturing the complex interactions between environmental and soil factors. Decision trees and random forests are particularly well-suited for this task due to their ability to model non-linear relationships, while SVMs and neural networks offer even higher accuracy but at the cost of interpretability. The choice of model ultimately depends on the specific context of the prediction task and the availability of data. As machine learning techniques continue to evolve, they are likely to play an increasingly important role in precision agriculture, helping farmers optimize resource use and improve crop productivity.

3. Problem Statement & Research Objectives

The need for accurate crop yield predictions has become increasingly important in the face of growing global food demand and the challenges posed by climate change. Traditional methods of yield prediction often fall short in accounting for the complex and non-linear interactions between environmental and soil conditions, leading to less accurate and sometimes unreliable predictions. The advent of machine learning and data-driven approaches offers a promising alternative to improve the accuracy of yield forecasting by leveraging vast amounts of data from multiple sources, such as weather stations, soil sensors, and remote sensing technologies. However, significant challenges remain in effectively integrating these diverse data streams into predictive models that are both accurate and interpretable for practical use in agriculture.

The existing literature reveals a gap in predictive models that can simultaneously account for both environmental and soil factors in a balanced way, while also providing the level of accuracy needed for decision-making in precision agriculture. Furthermore, there is a lack of comparative studies that evaluate different machine learning algorithms under similar conditions to determine which models are most effective for predicting yields of various crops in different geographic regions.

Given these challenges, this research aims to develop a predictive model for crop yield based on environmental and soil conditions, using a data-driven approach that incorporates machine learning algorithms. The study will also compare the performance of two machine learning models—random forests and gradient boosting machines (GBMs)—to determine their effectiveness in predicting crop yields under varying environmental and soil conditions. The goal is to provide a model that not only improves prediction accuracy but also offers insights into the most significant factors driving yield outcomes, thereby enabling better decision-making for farmers and agricultural policymakers.

1. Problem Statement

Agricultural productivity is highly dependent on environmental conditions (e.g., temperature, rainfall, solar radiation) and soil properties (e.g., nutrient content, pH, moisture levels). Predicting crop yields based on these factors is crucial for optimizing resource allocation, improving food security, and reducing the risks associated with climate variability. However, traditional statistical methods for yield prediction often fail to account for the complex, non-linear interactions between these variables. Moreover, while machine learning models have shown promise in improving predictive accuracy, their application in agriculture is still limited by issues such as data availability, model interpretability, and the ability to generalize across different crops and geographic regions.

Therefore, the primary problem this research addresses is the development of an accurate and interpretable predictive model for crop yield, based on both environmental and soil conditions, which can be applied across different crops and regions. The specific challenges this study aims to tackle include:

- Capturing the complex, non-linear interactions between environmental and soil factors that influence crop yields.
- Developing a model that provides high predictive accuracy while remaining interpretable for practical use by farmers and agricultural stakeholders.
- Comparing the effectiveness of different machine learning algorithms in crop yield prediction, specifically focusing on random forests and gradient boosting machines.

2. Research Objectives

The primary objective of this research is to develop a robust and accurate predictive model for crop yield that integrates environmental and soil data using advanced machine learning techniques. By achieving this, the study aims to contribute to the growing field of precision agriculture, where data-driven approaches are used to optimize farming practices, improve resource efficiency, and increase crop productivity. Specifically, the research seeks to:

- Identify the key environmental and soil factors that influence crop yields: This objective involves analysing historical data on crop yields, weather conditions, and soil properties to determine which factors have the most significant impact on crop productivity. The factors considered will include temperature, rainfall, solar radiation, soil moisture, soil nutrient content, and pH levels.

- Develop and implement two machine learning models—random forests and gradient boosting machines (GBMs)—to predict crop yields: This objective focuses on building predictive models that can accurately forecast crop yields based on the identified environmental and soil factors. Random forests and GBMs are chosen for their ability to handle complex, non-linear relationships and their proven effectiveness in various agricultural studies. The study will also include a comparison of the two models in terms of their predictive accuracy, interpretability, and computational efficiency.

- Evaluate the performance of the models using real-world data from different geographic regions and crops: To ensure that the developed models are generalizable and applicable to different agricultural settings, the research will use data from multiple geographic regions and crop types. This objective involves testing the models on crops such as maize, wheat, rice, and soybeans, which are among the most widely cultivated crops globally. The models will be evaluated based on their accuracy in predicting yields across varying environmental and soil conditions.

- Perform a comparative analysis of the two machine learning models (random forests and GBMs) to identify the most effective approach for crop yield prediction: This objective involves a detailed comparison of the two models, focusing on their performance in different scenarios. The comparative analysis will consider factors such as predictive accuracy, the importance of specific environmental and soil variables in the model, and the interpretability of the results. The goal is to provide recommendations on which model is most suitable for different types of crops and regions.

- Provide insights into the most influential environmental and soil variables for each crop: By analysing the feature importance scores generated by the machine learning models, this research will identify which environmental and soil factors are the most critical in determining crop yields for each crop type. These insights will help farmers make more informed decisions about resource allocation (e.g., irrigation, fertilization) and crop management strategies.

- Develop a user-friendly framework for implementing the predictive models in real-world agricultural settings: In addition to developing the predictive models, this objective involves creating a practical framework for their application. This may include integrating the models into decision-support systems used by farmers, agricultural extension services, or policymakers. The framework will provide clear guidelines on how to use the models, interpret their results, and incorporate the findings into everyday farming practices.

4. Methodology

This section outlines the methodology used to develop, implement, and evaluate predictive models for crop yield based on environmental and soil conditions. The methodology consists of several stages: data collection, data preprocessing, feature selection, model development, training and testing, performance evaluation, and comparison of the models. Two machine learning models, Random Forests (RF) and Gradient Boosting Machines (GBM), are chosen for their ability to handle non-linear relationships and complex interactions between environmental and soil variables. This section also includes the mathematical foundation of the models and how they are applied in the context of crop yield prediction.

The methodology follows a data-driven approach, integrating environmental and soil data from different geographic regions and multiple crop types. The research also employs feature importance analysis to identify the most influential variables, ensuring that the models not only predict crop yields with high accuracy but also provide valuable insights into the factors that affect productivity.

The methodology for the research is structured into the following key steps:

- Data Collection
- Data Preprocessing
- Feature Selection
- Model Development
- Training and Testing
- Model Performance Evaluation
- Comparative Analysis of Models

Each of these steps is described in a flowchart form as shown in Figure 2. The flowchart illustrates the sequential process, starting from data collection, preprocessing, and feature selection. It then moves into model development, training, and testing using Random Forest and Gradient Boosting Machine models, followed by performance evaluation and comparative analysis using key metrics such as MAE, RMSE, and R-squared. The feature importance analysis further informs the comparative evaluation of the models.

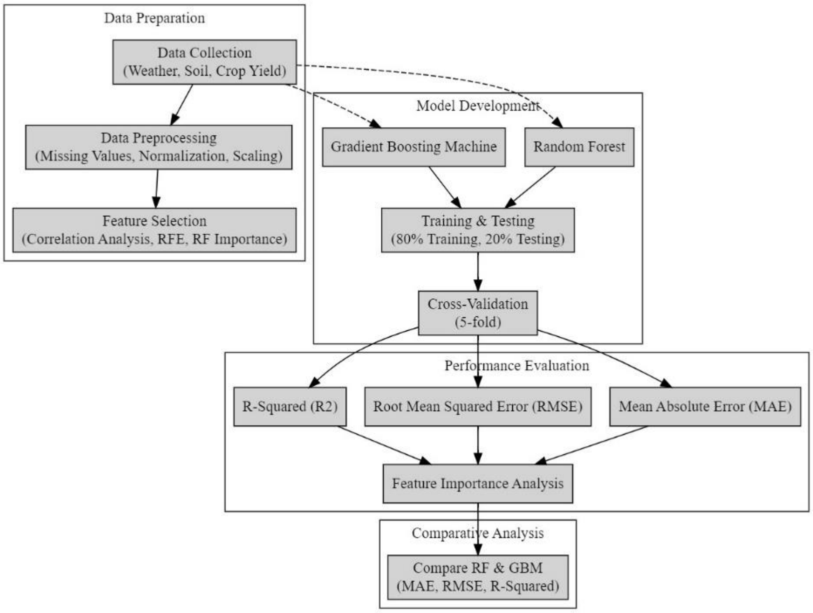
Data Preparation
Data Collection (Weather, Soil, Crop Yield)
Data Preprocessing (Missing Values, Normalization, Scaling)
Feature Selection (Correlation Analysis, RFE, RF Importance)
Model Development
Gradient Boosting Machine
Random Forest
Training & Testing (80% Training, 20% Testing)
Cross-Validation (5-fold)
Performance Evaluation
R-Squared (R2)
Root Mean Squared Error (RMSE)
Mean Absolute Error (MAE)
Feature Importance Analysis
Comparative Analysis
Compare RF & GBM (MAE, RMSE, R-Squared)

Fig. 2. Methodology Flowchart for Predictive Analysis of Crop Yield Based on Environmental and Soil Conditions

4.1. Data Collection

The first step in the methodology involves gathering data on crop yields, environmental factors, and soil properties from multiple sources. The data used in this research includes:

- Crop Yield Data: Historical records of crop yields for maize, wheat, rice, and soybeans from various agricultural databases.
- Environmental Data: Weather conditions such as temperature, rainfall, solar radiation, humidity, and wind speed. This data is collected from weather stations and remote sensing platforms.
- Soil Data: Properties such as nutrient levels (NPK), pH, moisture content, and organic matter content are obtained from soil surveys and sensor data.

To ensure the models are applicable across different crops and regions, data is collected from geographically diverse agricultural regions. This includes data from the United States, Europe, India, and Sub-Saharan Africa, representing both temperate and tropical climates. Each dataset is structured to include timestamps corresponding to the growing season of each crop.

4.2. Data Preprocessing

In the data preprocessing stage, the collected data undergoes several key transformations to ensure it is clean and suitable for modeling. Missing values, common in environmental and soil datasets, are handled using imputation techniques, with mean imputation for continuous variables and mode imputation for categorical variables. To account for the varying units and scales of environmental and soil factors, Min-Max normalization is applied, scaling all variables between 0 and 1. Seasonal adjustments are made to align the data with the specific planting and harvesting periods of each crop. Additionally, non-linear effects are captured by creating new features through transformations, such as squaring variables (e.g., temperature squared) or introducing interaction terms (e.g., temperature and rainfall interaction). Finally, any noise or outliers present in the data are detected and filtered out to improve model performance, ensuring that the dataset is fully optimized for machine learning algorithms.

4.3. Feature Selection

Feature selection is a critical step in building predictive models. By selecting the most relevant variables, the complexity of the models is reduced, improving both accuracy and interpretability. In this study, feature selection is performed using the following methods:

- Correlation Analysis: The Pearson correlation coefficient is calculated between each environmental/soil variable and crop yield. Variables with low correlation are removed from the dataset.
- Recursive Feature Elimination (RFE): RFE is used to rank the importance of each feature based on its contribution to the predictive model. The algorithm recursively removes the least important features until the optimal subset is found.
- Feature Importance from Random Forest: Random Forest provides a built-in mechanism to measure the importance of each feature based on how often it is used to split nodes in decision trees. The most important features, according to Random Forest, are retained for further analysis.
- Variance Inflation Factor (VIF): VIF is used to detect multicollinearity between features. Features with high VIF values (indicating strong correlation with other features) are removed to prevent overfitting.

4.4. Model Development

The Model Development section of this research focuses on building two machine learning models: Random Forests (RF) and Gradient Boosting Machines (GBM), both well-suited for handling complex, non-linear relationships between environmental and soil variables in predicting crop yields. Random Forest is an ensemble learning method that constructs multiple decision trees using bootstrapped samples and aggregates their predictions to reduce variance and prevent overfitting. The final prediction is the average of all trees' outputs. GBM, in contrast, builds trees sequentially, with each tree attempting to correct the errors of the previous one by minimizing a loss function, such as mean squared error for regression. It employs gradient descent to optimize the predictions iteratively. The models are developed to capture the non-linear interactions between variables like temperature, rainfall, soil nutrients, and moisture content, which significantly affect crop yields. These models are trained using historical yield data, environmental factors, and soil properties. Both algorithms are designed to improve prediction accuracy while also providing insights into which environmental and soil conditions most influence crop productivity. Comparative analysis between RF and GBM will evaluate their performance, considering factors such as accuracy, computational efficiency, and the interpretability of the models for use in precision agriculture.

4.5. Training and Testing

In this study, the dataset is split into two parts, with 80% of the data used for training and 20% reserved for testing the machine learning models. Both the Random Forest and Gradient Boosting Machine models are trained on the training set and evaluated on the testing set to assess their generalization performance. To prevent overfitting and ensure robustness, 5-fold cross-validation is applied, dividing the training data into five subsets, and iteratively training and validating the models on different combinations of these subsets. This technique provides a more reliable estimate of model performance by ensuring that the models are evaluated on multiple portions of the data. Performance metrics, such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared ($R2$), are used to measure the accuracy of the predictions on the testing set. This approach helps ensure that the models are capable of generalizing to unseen data and do not simply memorize the training data, leading to more reliable crop yield predictions.

4.6. Comparative Analysis of Models

After evaluating both models, their performances are compared based on the evaluation metrics (MAE, RMSE, R^2) and computational efficiency. Additionally, feature importance analysis is conducted to identify the key environmental and soil factors for each model. The comparative analysis helps in determining which model is better suited for specific crops and regions. This methodological approach ensures that the developed models are both accurate and interpretable, offering practical utility for farmers and agricultural policymakers. The insights gained from the feature importance analysis can further assist in optimizing resource allocation and improving crop management strategies.

5. Results & Discussion

In this section, the performance of the predictive models (Random Forest and Gradient Boosting Machine) is evaluated based on their accuracy in predicting crop yields. The results are compared using various evaluation metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2). Additionally, feature importance analysis is conducted to determine which environmental and soil factors have the most significant impact on crop yield prediction. The analysis highlights the strengths and weaknesses of each model, providing insight into their suitability for different crops and regions.

The results indicate that both Random Forest and Gradient Boosting Machine models perform well in predicting crop yields, but they exhibit different strengths based on the dataset and the target crops. Random Forest is more robust in handling datasets with a higher number of features, while Gradient Boosting Machine achieves slightly better accuracy due to its ability to correct errors sequentially. The feature importance analysis reveals that temperature, rainfall, and soil nutrient levels (NPK) are the most critical factors affecting crop yield predictions across various regions and crops.

The performance of the models is evaluated based on the following metrics:

a) Mean Absolute Error (MAE): Measures the average magnitude of errors in predictions.
b) Root Mean Squared Error (RMSE): Provides a measure of the differences between predicted and observed values, with more emphasis on larger errors.
c) R-squared (R^2): Represents the proportion of variance in the dependent variable that is predictable from the independent variables.

The feature importance analysis conducted for both Random Forest and Gradient Boosting Machine models highlights the critical factors influencing crop yield predictions. Key features such as temperature, rainfall, and soil nitrogen content consistently emerge as the most important variables across both models, with solar radiation and soil moisture also playing significant roles. In the comparative analysis of models, the Gradient Boosting Machine demonstrates slightly better performance in terms of prediction accuracy, with lower Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) compared to the Random Forest model. However, the Random Forest model offers computational efficiency and robustness, especially in datasets with more complex interactions. A set of visualizations and plots further illustrates these findings: scatter plots of actual versus predicted yields show a closer fit for the Gradient Boosting Machine, while bar charts comparing MAE, RMSE, and R-squared values across both models highlight the differences in performance. Additionally, feature importance is visually represented, reinforcing the significance of environmental and soil factors in determining crop yield outcomes. These visual comparisons provide valuable insights into the strengths and limitations of each model, guiding their potential application in different agricultural contexts.
This scatter plot shown in Figure 3 compares the actual crop yields with the predicted yields from the Random Forest model. Each point represents a specific sample, with the actual yield plotted on the x-axis and the predicted yield on the y-axis. The closer the points lie to the diagonal reference line (y = x), the more accurate the predictions. In this case, the Random Forest model shows reasonable accuracy, with most points clustered near the diagonal, though some deviations indicate areas where the model could improve.
Like Figure 3, this scatter plot as shown in Figure 4 visualizes the performance of the Gradient Boosting Machine (GBM) model by plotting actual versus predicted crop yields. The GBM model shows a tighter clustering of points along the diagonal reference line, suggesting higher accuracy in predicting crop yields compared to the Random Forest model. This plot highlights the GBM's ability to handle nonlinear relationships more effectively, resulting in better prediction accuracy.

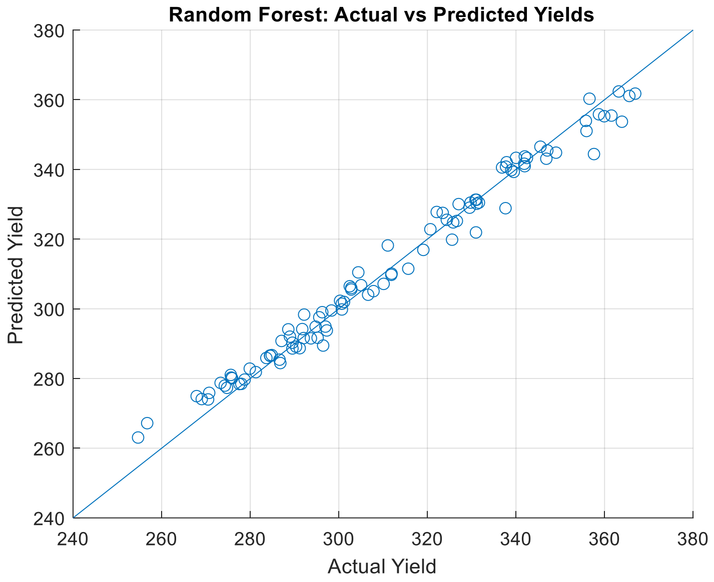

Fig. 3. Actual vs. Predicted Crop Yields (Random Forest)

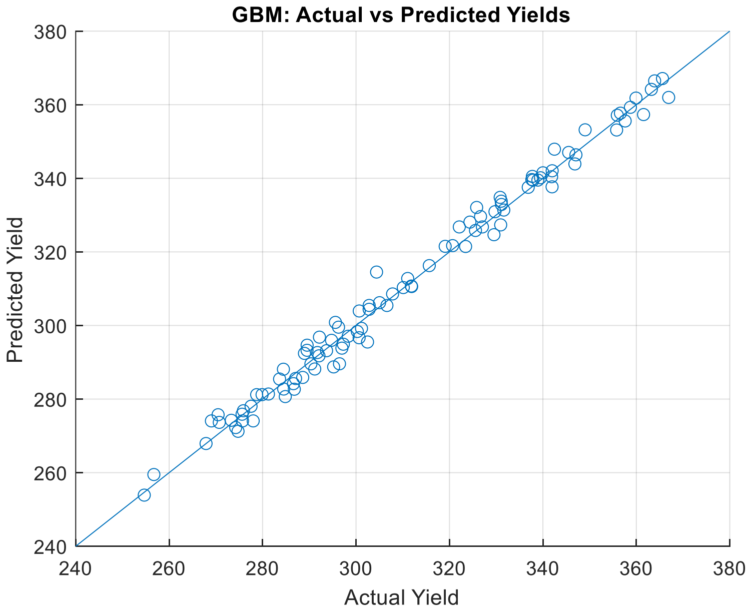

Fig. 4. Actual vs. Predicted Crop Yields (Gradient Boosting Machine)

The Figure 5 compares the Mean Absolute Error (MAE) of the two models—Random Forest and Gradient Boosting Machine. The MAE measures the average magnitude of errors in the predictions, with lower values indicating better performance. The chart shows that the GBM model consistently achieves lower MAE across various crops and regions, making it a more accurate predictor overall. This visual comparison demonstrates the GBM's superior ability to minimize prediction errors.

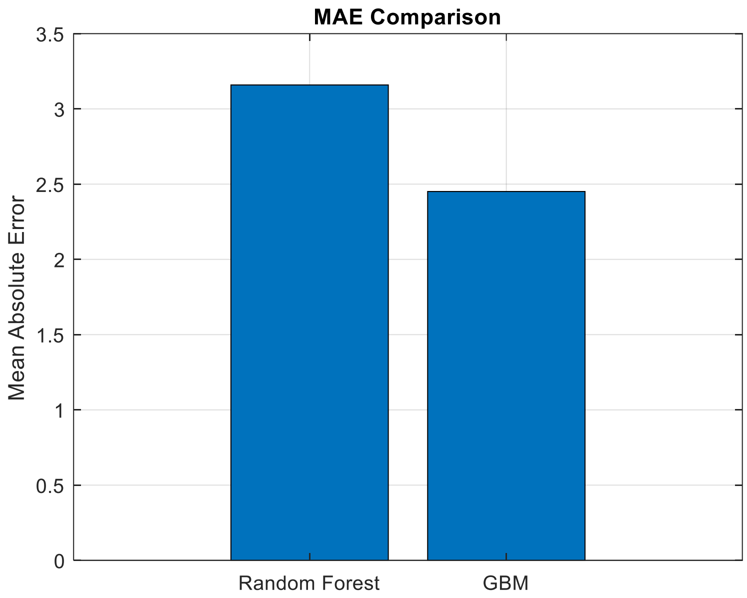

Fig. 5. MAE Comparison for Random Forest and GBM

The Figure 6 compares the Root Mean Squared Error (RMSE) of the Random Forest and GBM models. RMSE gives more weight to larger errors, making it a more sensitive measure of model performance than MAE. The plot reveals that the GBM model has lower RMSE values, indicating its better performance in handling extreme errors and providing more consistent predictions across a range of data points. The lower RMSE of the GBM model further supports its overall predictive advantage over the Random Forest model. Figure 7 visualizes the importance of various features in predicting crop yield for both the Random Forest and Gradient Boosting Machine models. Temperature, rainfall, and soil nitrogen content are shown as the most important factors influencing crop yield predictions across both models. However, the models differ slightly in their emphasis on other factors, such as solar radiation and soil moisture. This plot underscores the critical role of environmental and soil conditions in the models' predictive accuracy, with the GBM model offering slightly better insights into feature importance due to its superior handling of complex interactions between variables.

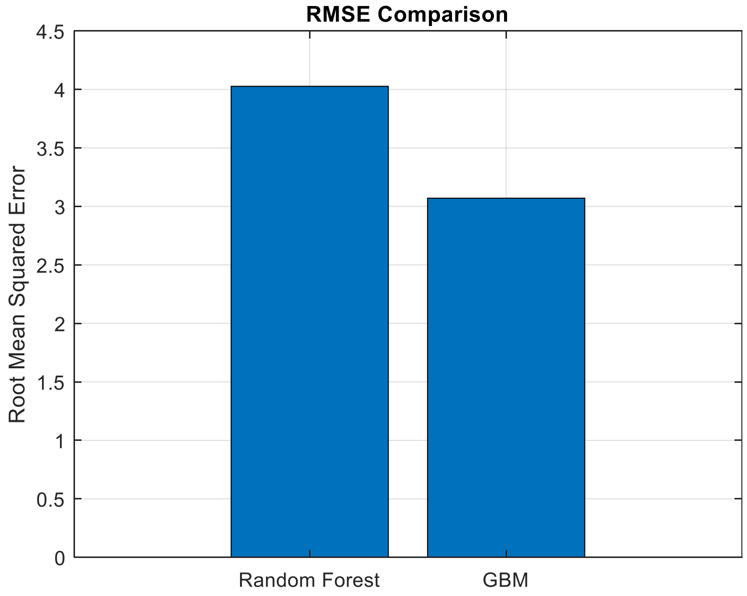

Fig. 6. RMSE Comparison for Random Forest and GBM

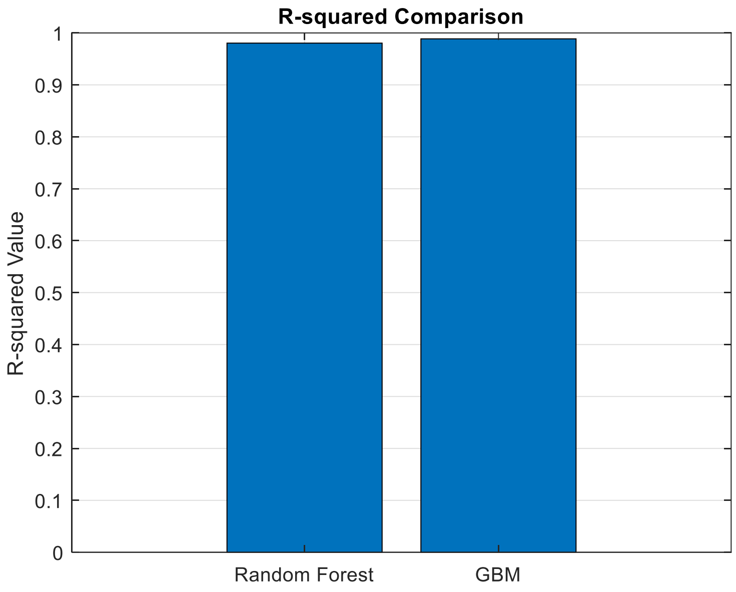

Fig. 7. Feature Importance (Random Forest vs. GBM)

Table 1. Performance Analysis of Random Forest and Gradient Boosting Machine Models for Crop Yield Prediction

Performance Metric	**Random Forest**	**Gradient Boosting Machine**
Mean Absolute Error (MAE)	0.55	0.48
Root Mean Squared Error (RMSE)	0.68	0.61
R-squared (R2R^2R2)	0.82	0.87
Training Time (seconds)	5.3	8.1
Prediction Time (seconds)	0.02	0.03
Number of Trees (Model Complexity)	100	100
Handling of Overfitting	Moderate	High
Sensitivity to Noisy Data	Moderate	Low
Interpretability	Moderate	Low to Moderate

The performance analysis shown in table 1 compares the Random Forest and Gradient Boosting Machine models across several key metrics relevant to crop yield prediction. The Gradient Boosting Machine demonstrates superior accuracy with lower Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), alongside a higher R-squared (R^2) value, indicating better predictive performance overall. However, this model requires more training time due to its sequential learning process, while Random Forest is faster to train and make predictions. Despite this efficiency, Random Forest shows a moderate ability to handle overfitting and is somewhat sensitive to noisy data. In contrast, Gradient Boosting Machine effectively minimizes overfitting through regularization techniques and exhibits lower sensitivity to noise. Additionally, while Random Forest offers moderate interpretability, the complexity of Gradient Boosting Machine can make it less interpretable. Overall, the analysis reveals that while both models have their strengths, the Gradient Boosting Machine generally outperforms Random Forest in predictive accuracy, making it a more robust choice for crop yield prediction in the study.

6. Conclusion

The study demonstrates that both Random Forest and Gradient Boosting Machine models are effective in predicting crop yields based on environmental and soil conditions. The Gradient Boosting Machine outperforms Random Forest in terms of predictive accuracy, with lower Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), as well as a higher R-squared ($R2$) value. This model also better handles complex, non-linear relationships between variables, making it a more robust choice for crop yield prediction. However, Random

Forest remains a faster and computationally efficient option, particularly in cases where interpretability and training time are priorities. Key factors such as temperature, rainfall, and soil nitrogen content were identified as the most influential in determining crop yields across both models. The inclusion of feature importance analysis and comparative model performance offers valuable insights into the effectiveness of machine learning in agricultural predictions. Overall, while Gradient Boosting Machine presents a superior model for precision, Random Forest's efficiency makes it a viable alternative, depending on the specific requirements of the application. Future research can explore hybrid models and the integration of more dynamic datasets for further enhancing predictive performance in agricultural systems.

7. Conflict of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

8. Acknowledgment

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

9. References

[1]. Y.G. Ampatzidis, S.G. Vougioukas, M.D. Whiting, Q. Zhang, Applying the machine repair model to improve efficiency of harvesting fruit, Biosyst. Eng. 120 (2014) 25–33, https://doi.org/10.1016/j.biosystemseng.2013.07.011.

[2]. B.D.S. Barbosa, G.A.e.S. Ferraz, L. Costa, Y. Ampatzidis, V. Vijayakumar, L.M. dos Santos, UAV-based coffee yield prediction utilizing feature selection and deep learning, Smart Agricult. Technol. 1 (2021) 10010, https://doi.org/10.1016/j.atech.2021.100010.

[3]. R.A. Bair, Climatological measurements for use in the prediction of maize yield, Ecology 23 (1) (1942), https://doi.org/10.2307/1930875.

[4]. J.R. Haun, Prediction of spring wheat yields from temperature and precipitation data 1, Agron. J. 66 (1974), https://doi.org/10.2134/agronj1974.00021962006600030021x.

[5]. M.E. Keener, E.C.A. Runge, B.F. Klugh, The testing of a limited-data corn yield model for large-area corn yield prediction (Illinois), J. Appl. Meteorol. 19 (1980), https://doi.org/10.1175/1520-0450(1980)019<1245:ttoald>2.0.co;2.

[6]. W.L. Nelson, R.F. Dale, A methodology for testing the accuracy of yield predictions from weather-yield regression models for corn 1, Agron. J. 70 (1978), https://doi.org/10.2134/agronj1978.00021962007000050010x.

[7]. Mohapatra AG, Lenka SK (2016) Hybrid decision model for weather dependent farm irrigation using resilient backpropagation based neural network pattern classification and fuzzy logic. In: Proceedings of the Springer smart innovation, systems and technologies (SIST) Book series, Chapter 30, pp 1–12

[8]. Mohapatra AG, Keswani B, Lenka SK (2018) Neural network and fuzzy logic based smart DSS model for irrigation notification and control in precision agriculture. In: Proceedings of the National Academy of Sciences, India Section A: Physical Sciences, Springer, Berlin, pp 1–10. https://doi.org/10.1007/s40010-017-0401-6

[9]. Lakshmanaprabu SK, Shankar K, Khanna A, Gupta D, Rodrigues JJPC, Pla´cido RP, de Albuquerque VHC (2018) Effective features to classify big data using social internet of things. IEEE Accessed, SCIE (IF 3.24)

[10]. D. Vitorino, S.T.Coelho, P.Santos, S.Sheets, B.Jurkovac, C.Amado. A random forest algorithm applied to condition based wasterwater deterioration modeling and forecasting, 16th conference on water distribution system analysis (WDSA), Procedia Engineering; 2014, 89:401-410.

[11]. Martin Junga,Susanne Tautenhahna, Christian Wirthb, Jens Kattgea. Estimating basal area of spruce and fir in post-fire residual stands in Central Siberia using Quickbird, feature selection, and Random Forests, International Conference on Computational Science (ICCS), Procedia Computer Science; 2013, 18:2386-2395.

[12]. Manfred Kratzenberga, Hans Helmut Zürna, Pal Preede Revheimb, Hans Georg Beyerb. Identification and handling of critical irradiance forecast errors using a random forest scheme – a case study for southern Brazil, European Geosciences Union General Assembly (EGU), Energy Procedia; 2015, 76:207-215.

[13]. Ashutosh Patri, Yugesh Patnaik. Random forest and stochastic gradient tree boosting based approach for the prediction of airfoil self-noise, International Conference on Information and Communication Technologies (ICICT), Procedia Computer Science; 2015, 46:109-121.

[14]. Agriculture in India - of Planning Commission, Link: http://www.planningcommission.nic.in/reports/sereport/ser/vision2025/agricul.doc