Image-Based Soil Quality Prediction using RGB Analysis and Machine Learning

G.Uday Kiran[1] , Srilakshmi .V[2], B.Moulika[3] and A.Rohith Reddy[4]

Department of Computer Science and Engineering (AI & ML)
B. V. Raju Institute of Technology, Narsapur, Telangana, India.
[1] udaykiran.goru@bvrit.ac.in, [2]srilakshmi.v@bvrit.ac.in, [3]22211a6617@bvrit.ac.in,
[4]22211a6608@bvrit.ac.in

Abstract. Traditional soil pH determination methods are time-consuming and labor-intensive, limiting their scalability for large-scale agricultural applications. The Image-Based Soil-Quality Prediction model addresses this by using RGB values from soil images to predict pH levels. Using a Random Forest Regressor, the model accurately predicts pH levels, while also diagnosing potential nutrient deficiencies by integrating soil nutrient data. This approach offers precise fertilizer and crop recommendations, enhancing soil health management. By combining advanced machine learning with image-based analysis, the model empowers farmers with actionable insights to optimize nutrient management, contributing to global food security and sustainable agricultural practices.

Keywords: soil pH, image-based prediction, RGB values, machine learning, nutrient deficiencies, sustainability.

1 Introduction

Agriculture is the cornerstone of many economies worldwide, and soil quality is a pivotal factor in agricultural productivity and sustainability [6]. Traditional soil analysis methods, although effective, can be costly and time-consuming. The integration of Machine Learning (ML) into soil quality prediction offers a promising alternative, providing efficient, scalable, and cost-effective solutions. ML can analyze vast datasets to uncover patterns and relationships among various soil parameters such as pH [1] , nutrient content[12], and moisture levels, which are crucial for determining soil health and suitability for specific crops.

The primary objectives of ML-based soil quality prediction projects include developing accurate predictive models, reducing soil testing costs, creating scalable solutions applicable across different regions, and aiding farmers in making informed decisions about crop selection[9] and soil management. These projects typically involve several key components: data collection [8] model development using various ML algorithms, model evaluation through metrics like accuracy and precision, and finally, deployment into user-friendly applications for real-time predictions.

Despite the potential benefits, these projects face challenges such as ensuring high-quality data, dealing with the variability in soil properties[10], and accounting for environmental factors that affect soil quality. However, the future looks promising with advancements in deep learning, ensemble learning, and the integration of IoT devices for real-time data collection. Collaborative efforts between agronomists, data scientists, and technologists are essential to enhance the accuracy and usability of ML models, ultimately leading to more sustainable and productive agricultural practices.

This project aims to create a cost-effective system for predicting soil pH[7] and recommending fertilizers[3][5] and crops[4][11] using image-based analysis and Machine Learning, specifically by training models like Random Forest Regressor on RGB features from soil images[2]. The goal is to provide a reliable tool for farmers to make informed decisions on soil management and crop production, promoting more efficient and sustainable agriculture.

2 Literature Review

The literature on soil pH prediction using digital imaging and machine learning (ML) highlights both the significant potential and the challenges of these innovative methods. While traditional soil pH measurement techniques, such as chemical assays and electrode-based methods, are highly accurate, they are often labour-intensive, time-consuming, and impractical for large-scale agriculture. These limitations have led to growing interest in digital imaging and ML as more efficient and scalable alternatives for managing soil nutrients.

Recent studies demonstrate the promise of digital imaging in soil property analysis. For instance, research by Konda Janardhan [1] and Ali Al-Naji et al. [2] shows that combining RGB imaging with ML techniques can effectively predict soil pH. Al-Naji et al. [2] used a standard video camera alongside a neural network to analyse soil colour variations, achieving high accuracy in predicting irrigation needs, thereby illustrating the potential of visual data in providing actionable insights into soil conditions. Similarly, Babalola et al. [7] employed RGB imaging to classify soil surface textures in uncontrolled field conditions, demonstrating the feasibility of image-based methods for soil characterization even in challenging environments. Machine learning models have further advanced soil pH prediction and the optimization of agricultural practices. For example, Prof. B.B. Vikhe et al. [3] showed how ML techniques could be used to forecast crop yields and recommend fertilizers, highlighting the effectiveness of these models in boosting agricultural productivity. Potnuru Sai Nishant et al. [9] explored various ML algorithms for crop yield prediction, showcasing their versatility and accuracy, and reflecting the broader trend of applying ML to solve complex agricultural problems.

Integrating agronomic principles with technological advancements is essential for developing comprehensive soil management solutions. Research by Jichong Han et al. [4] and Sultana et al. [5] underscores the importance of combining multi-source data with ML models to predict agricultural outcomes, emphasizing the need for models that incorporate not only soil parameters but also environmental factors for a more holistic understanding of soil health and crop suitability. Despite the progress, challenges remain, including issues related to data quality, the need for standardized protocols, and the integration of diverse data sources. Studies by Shailesh Kumar Dewangana et al. [6] and Dora Neina [10] highlight the importance of addressing these challenges by refining predictive models and understanding the impact of soil pH on overall soil health.

Novelty of the Current Work:
In contrast to existing studies, this research introduces a novel approach by integrating RGB analysis directly with pH prediction and nutrient recommendations using a Random Forest Regressor. While previous studies have demonstrated the effectiveness of machine learning in soil analysis, the current work emphasizes a comprehensive framework that not only predicts pH values but also provides tailored crop and fertilizer recommendations based on these predictions. Furthermore, the methodology employs a systematic RGB sampling technique, enhancing model robustness and reducing bias—elements that have not been fully explored in prior research. This holistic approach to soil quality assessment not only advances predictive accuracy but also offers practical solutions tailored for large-scale agricultural applications, thereby filling critical gaps in the existing literature.

In conclusion, the integration of digital imaging and ML marks a significant advancement in soil nutrient management, particularly for predicting soil pH. Although current methodologies represent substantial improvements over traditional techniques, ongoing research and development are crucial for overcoming existing limitations and enhancing the practical application of these approaches in large-scale agriculture. The continued evolution of these fields promises to deliver more precise and scalable solutions to soil management challenges, ultimately benefiting agricultural productivity and sustainability.

3 Datasets

The proposed project utilizes three main datasets to predict soil nutrients, recommend crops, and provide fertilizer suggestions based on soil pH values and RGB analysis.

The soilPH_rgb dataset contains 51 rows and is used for predicting soil pH based on the RGB values extracted from soil images. Each entry in this dataset includes a unique image identifier, the RGB values representing the color intensity of the soil, and the actual pH value, which serves as the target variable. Additionally, this dataset includes information on nutrient concentrations such as Nitrogen (N), Phosphorus (P), Potassium (K), and organic matter content, which are essential for determining soil health.

The Crop Recommendation dataset comprises 2201 rows and provides recommendations for suitable crops based on the predicted pH and nutrient levels of the soil. It links specific pH values and corresponding nutrient concentrations to crops that thrive in those conditions, helping farmers make informed crop selection decisions.

The Fertilizer Prediction dataset contains 100 rows and offers fertilizer recommendations by matching the required NPK (Nitrogen, Phosphorus, Potassium) values to those in the dataset. Based on the predicted pH and nutrient deficiencies, the system calculates the closest fertilizer match, ensuring optimal soil fertility management. These datasets collectively enable efficient soil quality assessment, crop recommendations, and fertilizer selection, improving agricultural productivity.

4 Methodology

The methodology for the Image-Based Soil-Quality Prediction Model is structured into several key phases: data collection, image processing, model training, pH prediction, nutrient retrieval, and fertilizer recommendation.

1. Data Collection
   The initial step involves gathering both image data and nutrient data. High-resolution images of soil samples are captured under controlled lighting conditions and at a consistent camera angle to ensure uniformity across all samples. These images are stored in a standardized format (e.g., .webp) to facilitate processing and reduce file size without sacrificing quality. Additionally, datasets from soil testing laboratories are compiled, including:
   - soilPH_rgb dataset (RGB values linked to corresponding pH measurements).
   - crop recommendation dataset (detailing the nitrogen (N), phosphorus (P), and potassium (K) requirements for various crops).
   - fertilizer dataset (listing available fertilizers along with their nutrient compositions).

2. Image Processing
   In this phase, the soil images are processed using image libraries like Python Imaging Library (PIL). Each image is opened and converted to RGB mode, ensuring consistent color representation across all samples. The resolution of the image and the environment's lighting intensity during image capture are defined as simulation parameters to minimize variability.

   RGB values are extracted from each pixel in the image, creating a comprehensive dataset of color information. To enhance model robustness, random sampling of RGB values is performed. This step is critical for reducing bias and ensuring that a representative subset of the RGB values is used for model training.

3. Model Training
   The extracted RGB data, along with the corresponding pH levels, are loaded from the CSV file into a structured format for analysis. The dataset is split into training (80%) and testing (20%) sets, allowing for efficient model evaluation. The key simulation parameters for the machine learning phase include:

- Hyperparameters for the Random Forest Regressor, such as the number of trees (e.g., 100), the maximum depth of each tree, and the minimum samples per leaf.
- Performance Metrics: The Mean Squared Error (MSE) is used as the key performance metric to evaluate prediction accuracy.

A Random Forest Regressor is then trained to model the relationship between RGB values and pH levels. During this phase, the model is fine-tuned using cross-validation techniques and hyperparameter optimization to ensure accurate predictions.

4. pH Value Prediction
   Once the model is trained, it is used to predict the soil's pH value based on the randomly selected RGB values. The simulation parameters governing this phase include the thresholds for RGB-to-pH mapping and the acceptable error margin for pH predictions. This pH prediction informs the subsequent nutrient analysis.

5. Nutrient Retrieval
   After the pH value is predicted, the model retrieves nitrogen (N), phosphorus (P), and potassium (K) values associated with the predicted pH level by querying the crop recommendation dataset. The pH ranges used to map crops and their corresponding NPK values are set as key simulation parameters. This ensures precise matching between the predicted pH and the nutrient needs of various crops, enabling effective nutrient deficiency diagnostics.

6. Fertilizer Recommendation
   In the final phase, a Euclidean distance-based approach is implemented to recommend the most appropriate fertilizer. The predicted NPK values from the nutrient retrieval step are compared with the nutrient compositions in the fertilizer dataset. The Euclidean distance formula serves as a simulation parameter to ensure that the recommended fertilizer has the closest match to the soil's nutrient needs. By minimizing this distance, the system identifies the best fertilizer for the soil, thereby optimizing soil fertility and crop yields.

The integration of simulation parameters throughout the methodology ensures consistency and accuracy in the prediction process:

- Image Acquisition: Resolution, camera angle, and lighting intensity.
- Image Processing: RGB mode conversion and random sampling of pixel values.
- Machine Learning Model: Random Forest hyperparameters (e.g., number of trees, maximum depth), data splitting ratios, and MSE as the evaluation metric.
- pH Prediction: Thresholds for RGB-pH mapping and acceptable error margins.

- Nutrient and Fertilizer Matching: pH ranges for nutrient retrieval, Euclidean distance for fertilizer recommendation.

This comprehensive methodology integrates advanced image processing with machine learning techniques, providing an efficient framework for predicting soil pH and recommending appropriate fertilizers and crops. By leveraging technology, the model aims to enhance soil health management, empower farmers with actionable insights, and contribute to sustainable agricultural practices.

4.1 Random Forest

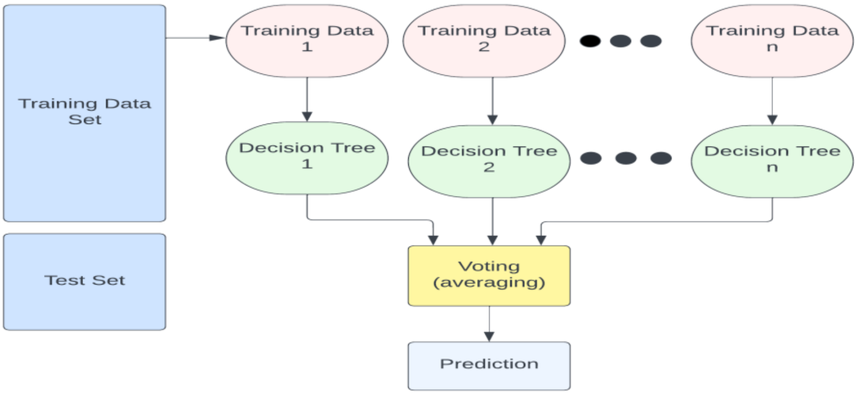


Figure 1 Architecture of Random Forest

The image illustrates the functioning of the Random Forest algorithm, which operates by splitting the training dataset into multiple subsets and training separate decision trees on each subset. Each tree independently learns patterns from different portions of the data. In regression tasks, the final prediction is made by averaging the predictions from all the individual trees. This ensemble method enhances model robustness by reducing the risk of overfitting and ensuring more accurate predictions. By aggregating the output of multiple decision trees, the Random Forest algorithm provides a reliable and efficient approach to handling complex data relationships, making it effective for predicting continuous values such as soil pH based on RGB values from images.

4.2 Soil Quality Prediction

Image-Based Soil Quality Prediction using RGB Analysis and Machine Learning focuses on utilizing RGB images of soil samples to predict soil parameters such as pH and nutrient content, alongside providing crop and fertilizer recommendations. The system processes soil images, analyzes their RGB values, and uses these values to predict crucial soil characteristics, essential for enhancing agricultural productivity. Through the application of Machine Learning models, specifically Random Forest, the system automates soil quality assessment and delivers actionable insights for optimizing crop yields and soil health.

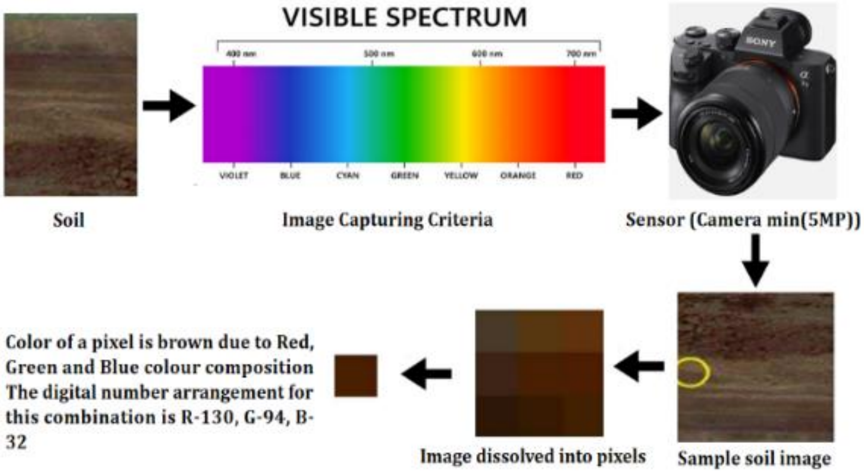

VISIBLE SPECTRUM
400 nm
500 nm
600 nm
700 nm
VIOLET
BLUE
CYAN
GREEN
YELLOW
ORANGE
RED
Soil
Image Capturing Criteria
Sensor (Camera min(5MP))
Color of a pixel is brown due to Red, Green and Blue colour composition The digital number arrangement for this combination is R-130, G-94, B-32
Image dissolved into pixels
Sample soil image

Figure 2 Methodology of RGB Analysis for Soil quality

The first step involves image acquisition and processing. Digital images of soil samples are captured using a camera or smartphone and converted to RGB mode using the Python Imaging Library (PIL), ensuring consistent color representation. RGB values are then extracted from each pixel in the image, where the intensities of Red (R), Green (G), and Blue (B) channels describe the pixel’s color. This data serves as the input for further analysis.

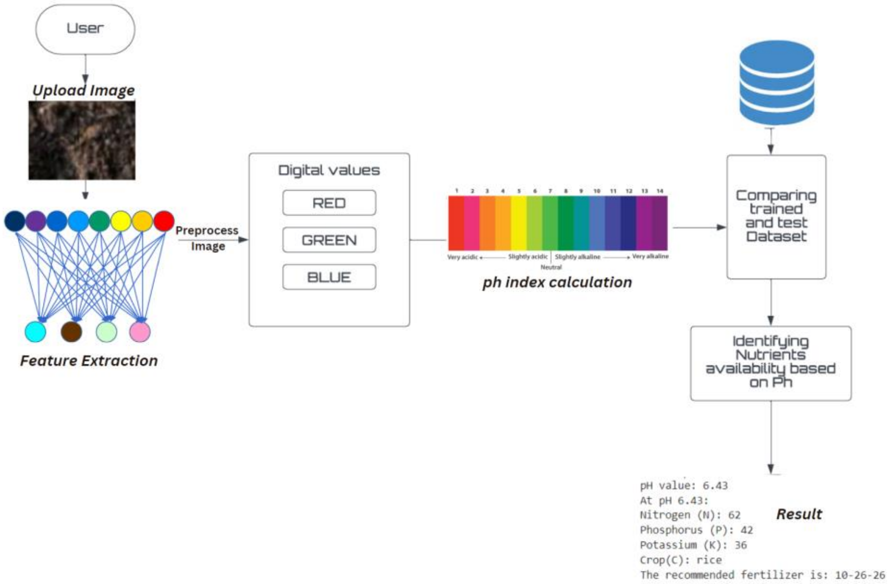
Afterward, random sampling selects a subset of the extracted RGB values to represent the soil sample's overall color characteristics. This subset is fed into a Machine Learning model for training. A Random Forest Regressor is used, trained on a dataset of RGB values and their corresponding pH levels. The Random Forest algorithm creates multiple decision trees and averages their outputs to ensure accurate predictions of soil pH based on RGB inputs.

The system then predicts the pH value of new soil samples using the trained model. With the predicted pH value, corresponding Nitrogen (N), Phosphorus (P), and Potassium (K) levels are retrieved from a crop recommendation dataset. This dataset outlines the ideal NPK ratios for different pH levels and provides suggestions for crops that are best suited to the predicted pH conditions.

In addition to crop recommendations, the system offers fertilizer recommendations by calculating the Euclidean distance between the required NPK values and those in a fertilizer dataset. The fertilizer with the closest match to the required NPK values is recommended as the best option for optimizing soil fertility.

By combining RGB analysis with Machine Learning, Image-Based Soil Quality Prediction using RGB Analysis and Machine Learning enhances soil management practices. The integration of computer vision and predictive modeling enables accurate soil quality assessment, efficient fertilizer use, and informed crop selection, ultimately improving agricultural productivity and sustainability.

4.3 Architecture of Proposed Model



```
pH value: 6.43
At pH 6.43:
Nitrogen (N): 62
Phosphorus (P): 42
Potassium (K): 36
Crop(C): rice
The recommended fertilizer is: 10-26-26
```

Figure 3 Architecture of Soil Quality Prediction using RGB

The architecture begins with the user uploading a soil image, which is preprocessed and analyzed to extract RGB values. A machine learning model, specifically a Random Forest Regressor, is trained to predict the soil's pH based on these RGB values. The predicted pH value is then used to query a crop recommendation dataset, retrieving the optimal nitrogen (N), phosphorus (P), and potassium (K) values for crops suited to that pH level. The system further compares these NPK values with those found in a fertilizer dataset, using Euclidean distance to determine the best matching fertilizer. Finally, it outputs the predicted pH value, the recommended NPK levels, the suitable crop, and the most appropriate fertilizer for the soil condition.

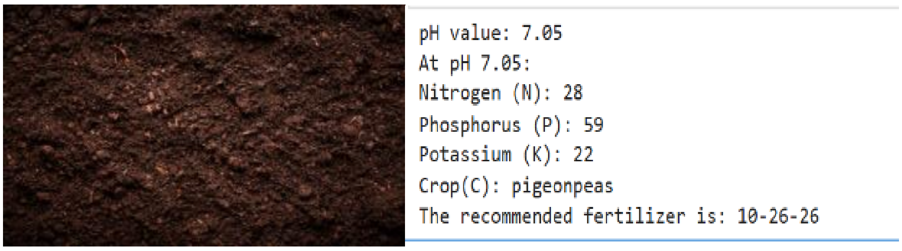
4.4 Summary

The Image-Based Soil Quality Prediction using RGB Analysis and Machine Learning project focuses on predicting soil characteristics like pH and nutrient content using RGB images of soil samples. The system processes these images to extract RGB values, which are then analysed using a Machine Learning model, specifically a Random Forest Regressor, to predict soil pH. Based on the predicted pH, the system retrieves corresponding nutrient levels (N, P, K), crop recommendations, and fertilizer suggestions. The project leverages three datasets: one for soil pH and RGB values, another for crop recommendations based on pH and nutrient levels, and a third for fertilizer recommendations. This approach automates soil quality assessment, aiding in efficient fertilizer use and optimal crop selection, ultimately enhancing agricultural productivity and sustainability.

5 Results and Discussion

The image-based soil quality prediction model integrates machine learning and image processing techniques to accurately predict soil pH and recommend suitable fertilizers. High-resolution images of soil samples are processed to extract RGB values, which are then utilized by a trained Random Forest Regressor to estimate the soil pH. The model is evaluated using Mean Squared Error (MSE) as the primary performance metric, which reflects the accuracy of the pH predictions. Following the pH estimation, nitrogen (N), phosphorus (P), and potassium (K) values are retrieved from a crop recommendation dataset, enabling the diagnosis of potential nutrient deficiencies. Subsequently, a Euclidean distance-based approach is employed to match the retrieved NPK values with available fertilizers, ensuring the identification of the most appropriate option. This comprehensive methodology facilitates precise pH prediction, effective nutrient analysis, and tailored fertilizer recommendations, ultimately supporting enhanced soil health and sustainable agricultural practices.

The results of the model are quantified through several key metrics. The Mean Squared Error (MSE) indicates the average squared difference between predicted and actual pH values, with lower values signifying better model performance. The Mean Absolute Error (MAE) provides insight into the average absolute discrepancies between predictions and actual values, also with lower values indicating improved fit. However, the R-squared (R^2) value of 0.0829 suggests that the model explains only about 8.29% of the variability in pH levels, pointing to limitations in predictive power. The predicted pH value derived from the RGB data, alongside the specific nutrient requirements for optimal crop growth, underpins the tailored fertilizer recommendation, which, in this instance, is Urea. This dual focus on prediction accuracy and practical applicability underscores the model's potential impact on soil health management in agricultural settings.



```
pH value: 7.05
At pH 7.05:
Nitrogen (N): 28
Phosphorus (P): 59
Potassium (K): 22
Crop(C): pigeonpeas
The recommended fertilizer is: 10-26-26
```

Figure 4 Input image and its respective output

Table 1. Regression Metrics and Results Table

Metric	Value
Mean Squared Error (MSE)	0.3583
Mean Absolute Error (MAE)	0.5570
R-squared (R^2)	0.0829
Predicted pH value	6.8653
Nitrogen (N)	76
Phosphorus (P)	39
Potassium (K)	24
Crop (C)	Maize
Recommended Fertilizer	Urea

6 Conclusion

Based on the analysis and comparison of different classification approaches, it is evident that the accuracy scores of classifiers, such as Random Forest, highlight their effectiveness, with the ensemble method often outperforming others for the given dataset. Model selection depends on task-specific requirements and dataset characteristics, where Logistic Regression may suit linearly separable data, while Decision Trees and Random Forests handle more complex decision boundaries. Future improvements could involve exploring advanced feature engineering techniques, fine-tuning hyperparameters through grid or random search, and investigating ensemble methods like GBM, AdaBoost, or XGBoost for enhanced performance. Additionally, experimenting with deep learning models, particularly for complex datasets with high-dimensional features, could yield better results. Emphasizing model interpretability and performing cross-validation will also ensure more robust estimates of performance and generalization to unseen data, ultimately refining classification models for real-world applications.

References

1. Janardhan, Konda. "Determination of pH in Soil Using Deep Learning and Digital Image Processing." International Research Journal of Innovations in Engineering and Technology 4.3 (2020): 66.
2. Al-Naji, A., et al. "Soil color analysis based on a RGB camera and an artificial neural network towards smart irrigation: A pilot study. Heliyon, 7 (1), e06078." (2021).
3. Bondre, Devdatta A., and Santosh Mahagaonkar. "Prediction of crop yield and fertilizer recommendation using machine learning algorithms." International Journal of Engineering Applied Sciences and Technology 4.5 (2019): 371-376.
4. Han, Jichong, et al. "Prediction of winter wheat yield based on multi-source data and machine learning in China." Remote Sensing 12.2 (2020): 236.

5. Sultana, J., M. N. A. Siddique, and M. R. Abdullah. "Fertilizer recommendation for agriculture: practice, practicalities and adaptation in Bangladesh and Netherlands." International Journal of Business, Management and Social Research 1.1 (2015): 21-40.
6. Dewangan, S. K., et al. "The effects of soil ph on soil health and environmental sustainability: A review." JETIR, Jun (2023).
7. Babalola, Ekunayo-oluwabami, Muhammad H. Asad, and Abdul Bais. "Soil surface texture classification using RGB images acquired under uncontrolled field conditions." IEEE Access (2023).
8. Kainth, Kamaljeet, Baljit Singh, and Virender Singh. "A REAL TIME COMPARATIVE ANALYSIS OF RGB IMAGES CAPTURED FROM VARIOUS DSLR CAMERAS."
9. Nishant, Potnuru Sai, et al. "Crop yield prediction based on Indian agriculture using machine learning." 2020 international conference for emerging technology (INCET). IEEE, 2020.
10. Neina, Dora. "The role of soil pH in plant nutrition and soil remediation." Applied and environmental soil science 2019.1 (2019): 5794869.
11. Dighe, Deepti, et al. "Survey of crop recommendation systems." IRJET 5 (2018): 476-481.
12. Ahmed, Usman, et al. "A nutrient recommendation system for soil fertilization based on evolutionary computation." Computers and Electronics in Agriculture 189 (2021): 106407.