WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Soil Classification and Crop Prediction Using Machine Learning

Prajakta Palaskar[1], Vaishnavi Padyal[2], Mrunalini Wagh[3], Prof. Jayashree Jadhav[4], Shruti Yele[5]

[1,2,3,4,5]Department of Computer Engineering Bharati Vidyapeeth College of Engineering for Women. Pune, India.
[1]palaskarprajakta8@gmail.com, [2] vaishnavipadyalbvcoew@gmail.com, [3] mrunalini184@gmail.com, [4] jayashree.jadhav@bharatividyapeeth.edu, [5]shrutiyele2002@gmail.com

ABSTRACT –

With the growing demand for increased efficiency in agriculture, modern farming is rapidly embracing advanced technologies to improve crop yield and resource management. This paper presents a smart, machine learning–based solution for soil classification and crop prediction, leveraging the strengths of Convolutional Neural Networks (CNNs) and Support Vector Machines (SVMs). The system is designed to analyze both soil imagery and related environmental parameters to accurately identify soil types and suggest suitable crops. Additionally, it offers tailored recommendations for fertilizers and herbicides to further optimize agricultural output. CNNs are utilized for extracting meaningful features from soil images, while SVMs handle classification tasks based on both visual and numerical inputs. The ultimate objective is to support farmers and agronomists in making precise, data-driven decisions that enhance productivity and sustainability. Looking ahead, the system can be extended with IoT integration, mobile access, and multilingual interfaces to increase its accessibility and applicability on a national scale.

Keywords: *Predictive Analytics, Crop Yield Optimization, Soil Management, Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs), Precision Agriculture, Resource Efficiency, Image based Assessment*

I. Introduction

Agriculture is a cornerstone of economic stability and a key livelihood source, especially in developing countries where many people depend on farming for their income. However, despite its importance, the agricultural sector still grapples with issues like low crop yields, inefficient resource use, and poor decision-making regarding crop selection. One of the core challenges behind these problems is the lack of reliable information about soil health and its suitability for different crops. Many farmers continue to rely on age-old practices or trial-and-error methods, which don't always reflect the actual needs of the soil or the prevailing environmental conditions. As a result, productivity suffers, and natural resources are often wasted.

To address these issues, this project introduces a machine learning–based system designed to automate and improve how soil is assessed and how crops are chosen. The goal is to support farmers and agricultural planners in making smarter, data-driven decisions that lead to better yields and more sustainable farming. The system is built around two main components: soil classification and crop recommendation.

First, the system analyzes various chemical and physical properties of soil—such as pH, moisture, and the levels of nitrogen, phosphorus, and potassium—to categorize it into different types. Using this data, a machine learning model is trained to identify soil categories accurately.

Once the soil is classified, the system shifts to crop recommendation. It does this by using past agricultural data that includes soil characteristics, climate factors, and crop performance. The model learns from this information to predict which crops are most likely to grow well under the given conditions. This method not only boosts yield but also ensures smarter use of resources like water and fertilizer.

What makes this system particularly valuable is its ability to remove much of the uncertainty in farming. It provides clear, science-backed recommendations that farmers can use confidently. The platform is also designed to be easy to access, either through a website or a mobile app, so that even those with limited technical knowledge can benefit from it.

In essence, this project brings modern machine learning techniques into the world of traditional agriculture. It aims to close the information gap, improve farming outcomes, and encourage more sustainable practices. By offering practical, accessible tools, this approach has the potential to reshape how farming decisions are made and significantly uplift agricultural productivity and rural livelihoods.

II. LITERATURE SURVEY

In 2020, Dr. Y. Jeevan Nagendra Kumar and colleagues investigated the use of supervised machine learning algorithms for crop yield prediction in the agriculture sector. Their research demonstrated that models such as Random Forest are particularly effective in predicting yields by analyzing features

like soil nutrients, weather patterns, and crop types. Random Forest performed best due to its ability to handle complex, high-dimensional data and reduce overfitting. The study emphasized the importance of feature selection and data preprocessing for improving accuracy. However, the study also highlighted some limitations, such as the dependency on high-quality historical agricultural data, regional variations affecting model transferability, and the difficulty of predicting dynamic environmental factors. Additionally, the interpretability of ensemble models like Random Forest posed a challenge for practical understanding by farmers. Despite these challenges, the research supports the use of machine learning in advancing data-driven decision-making in agriculture [1].

In another 2020 study, P. R. Dhumal and colleagues explored soil classification using machine learning techniques such as Random Forest, Decision Tree, and SVM. The models were trained on soil parameters like pH, nitrogen, phosphorus, and potassium, and Random Forest yielded the highest classification accuracy. The findings underscored the potential of ML for reliable and efficient soil type categorization to assist in crop planning. Nevertheless, challenges such as the availability of clean, labeled datasets, performance degradation on unseen or rare soil types, and limited model interpretability for non-technical users were noted. Furthermore, the absence of real-time soil monitoring integration was identified as a limitation for practical application [2].

Meenakshi B. Shewale et al. in 2020 proposed a combined system for soil classification and crop suggestion using machine learning models including Random Forest, Decision Tree, and SVM. By using soil characteristics such as pH, nitrogen, phosphorus, and potassium levels, the system classified soil types and recommended suitable crops accordingly. This integrated approach helped improve decision-making for farmers and demonstrated potential for enhancing agricultural productivity. The study emphasized the role of data preprocessing and feature selection in achieving higher accuracy. However, it also faced limitations such as reliance on high-quality soil data, poor performance for unrepresented crops or soil types, lack of real-time updates, and challenges in usability for non-experts [3].

In 2017, P. S. Vijayabaskar and colleagues focused on crop prediction using predictive analytics methods such as Regression, in 2017, P. S. Vijayabaskar and colleagues focused on crop prediction using predictive analytics methods such as Regression, Decision Trees, and Random Forests. Their models effectively predicted crop yields using historical data on weather, soil quality, and farming practices. The study showed that combining multiple data sources improved model accuracy and robustness. These predictive tools enabled farmers to make proactive decisions, mitigating risks related to crop failure and market uncertainty. However, limitations included dependence on large volumes of diverse data, limited ability to anticipate environmental disruptions, the need for significant computational resources, and the lack of real-time adaptability. Additionally, regional differences required continual model retraining for best performance. [4].

In 2021, Z. M. Shaikh, S. S. Gaikwad, and P. R. Tadas carried out a study to predict suitable crops using machine learning techniques, with a focus on the Random Forest algorithm. Their model was trained on a dataset that included various agricultural factors such as nitrogen, phosphorus, potassium content, temperature, humidity, soil pH, and rainfall levels. To assess the effectiveness of their approach, they compared Random Forest with other popular algorithms like Naive Bayes, Decision Tree, and Logistic Regression. The results showed that Random Forest delivered better accuracy in predicting the right crops. Despite this success, the study was limited to a static dataset, meaning it lacked the ability to adapt to real-time environmental changes. The authors noted that including live weather updates and soil conditions in future models could significantly improve prediction outcomes. They also suggested that integrating Internet of Things (IoT) devices could help automate data collection, making the system more dynamic and useful for real-time agricultural decision-making [5].

III. RESEARCH GAP

In today's world, technology is playing a big role in improving agriculture, but many farmers still depend on traditional ways to identify soil types and decide which crops to grow. These methods are often based on personal experience or guesswork, which can lead to lower productivity. While some tools and models do exist, they usually require expensive equipment, large datasets, or technical knowledge—things that aren't always available to farmers in rural or less-developed areas. That's where our project comes in.

We've built a system that helps classify soil types using images and suggests the most suitable crops by looking at different factors like the levels of nitrogen, phosphorus, and potassium in the soil, as well as the temperature, humidity, pH, and rainfall. What makes our project special is that it's simple to use and works through a web interface. A user can just upload a soil image and enter the basic soil details to get an instant crop recommendation.

Although the accuracy of our model is still improving due to limited data, it shows a lot of potential. In the future, we aim to add more soil images and data to make the system even better. We're also planning to include features that explain how the predictions are made, so users can understand and trust the results. Overall, this project is a step toward creating easy-to-use, low-cost tools that can help farmers make smarter and more informed decisions.

IV. METHODOLGY

A. Data Set

For our project, we've gathered two helpful datasets from Kaggle—one that contains images of different types of soil, and another that provides detailed soil and weather data linked to suitable crops. The first dataset, called the Soil Image Dataset, includes photographs of various soil types such as red, black, alluvial, sandy, and clayey soil. Each image is neatly organized in folders based on its category, making it easier to use for training models that

can visually identify soil types. This kind of data is especially useful when we want to build systems that can recognize soil just by taking a picture—something that could be really helpful for farmers or agriculture workers in remote areas without access to lab testing.

The second dataset, known as the Crop Recommendation Dataset, is in a table format and includes important details like the levels of nitrogen, phosphorus, and potassium in the soil, as well as environmental factors such as temperature, humidity, pH, and rainfall. Every row in the dataset ends with a suggested crop that fits those specific conditions for example, rice, maize, or cotton. This data can be used to train machine learning models that predict which crop is most likely to grow well based on the soil's nutrients and the climate.

Together, these two datasets give us a complete view—one that combines how the soil looks with what's actually in it. By using both image-based and numeric data, we can create smarter systems that help make farming decisions more accurate, practical, and accessible to everyone, especially in areas where expert advice is hard is hard to find.

B. Model Structure

The proposed system for Soil Classification and Crop Prediction integrates Convolutional Neural Networks (CNN) with structured environmental data, utilizing a hybrid architecture aimed at accurate classification of soil types and crop recommendation. The design leverages both image-based and numerical feature inputs to form a comprehensive decision-making pipeline suitable for agricultural planning.

The image-processing component employs a CNN framework beginning with an input layer that accepts RGB satellite images of size 224×224×3, capturing spatial land characteristics such as vegetation texture, field boundaries, and terrain variations. The first layer of the network consists of a 2D convolutional layer with 32 filters and a 3×3 kernel size, followed by a ReLU activation function to introduce non-linearity. This configuration enables the model to detect low-level features like edges, textures, and color contrasts present in the land cover.

Subsequently, a MaxPooling layer with a 2×2 window is applied to down sample the feature maps, reducing dimensionality and computational load while preserving essential features. The resultant feature maps are then flattened into a one-dimensional vector, which is passed to a Dense layer with 128 units and ReLU activation. This fully connected layer allows for deeper abstraction and pattern recognition from the spatial features extracted by earlier convolutional stages.

In parallel, the system accepts structured soil data comprising key agricultural indicators such as pH, moisture content, and NPK levels. These numerical features undergo a normalization and scaling process to standardize their range, ensuring compatibility with the flattened CNN features.

Following feature extraction from both the image and structured data paths, the two streams are concatenated and re-normalized, resulting in a unified feature representation that is fed into a Support Vector Machine (SVM) classifier. The SVM was chosen for its effectiveness in high-dimensional spaces and ability to construct clear decision boundaries for multi-class classification.

The final output layer of the system predicts two key agricultural insights: the type of soil (e.g., black, alluvial, sandy) and the recommended crop, based on a combination of visual land attributes and soil chemistry data.

For model training, the CNN layers were initialized with random weights and optimized using Adam optimizer with an appropriate learning rate to balance convergence speed and stability. The classification loss was computed using categorical cross-entropy, and performance was evaluated using accuracy as the primary metric.

This hybrid model design supports precision agriculture, offering an end-to-end intelligent system that harnesses both visual and environmental data to guide crop selection and land management strategies.

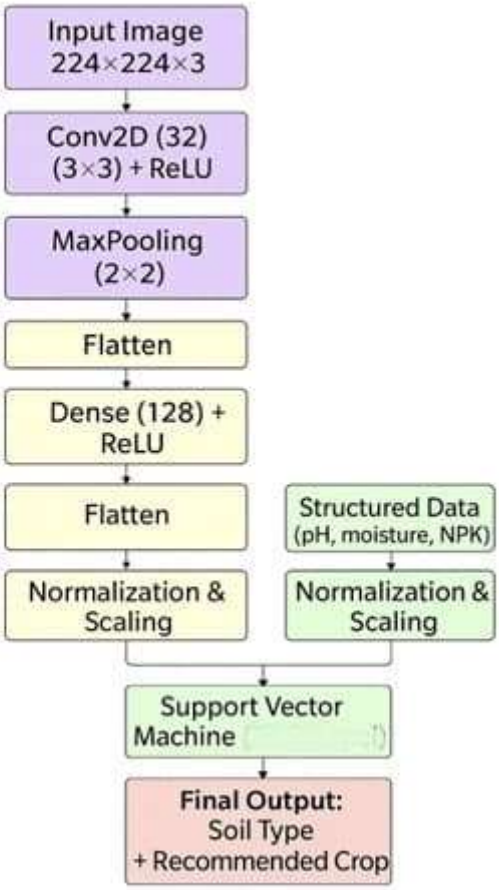

Fig. 1 Model structure

Following are the images of soil types:

A. Red Soil

- Reddish-brown color (due to iron oxide)
- Dry and porous texture
- May appear loose and crumbly

B. Black Soil

- Deep black or dark gray color (rich in organic matter and clay)
- Moist and sticky when wet
- May appear cracked when dry


C. Alluvial soil

- Light grey to ash brown color (depending on source and depth)
- Soft, fertile, and loamy texture
- May appear layered or slightly sandy in areas


D. Sandy Soil

- Light brown to reddish color (depending on minerals present)
- Loose and gritty texture
- May appear dry and non-cohesive

C. Model Implementation

The implementation of the proposed model for Soil Classification and Crop Prediction integrates both image-based and textual data processing using a hybrid machine learning architecture. The system accepts soil images and corresponding textual soil data as inputs, both sourced from a centralized database. All soil images are pre-processed by resizing and normalization to ensure consistency in dimensions and intensity levels. This standardization facilitates stable training of the image processing pipeline and ensures robustness to variability in image acquisition conditions, such as lighting or resolution.

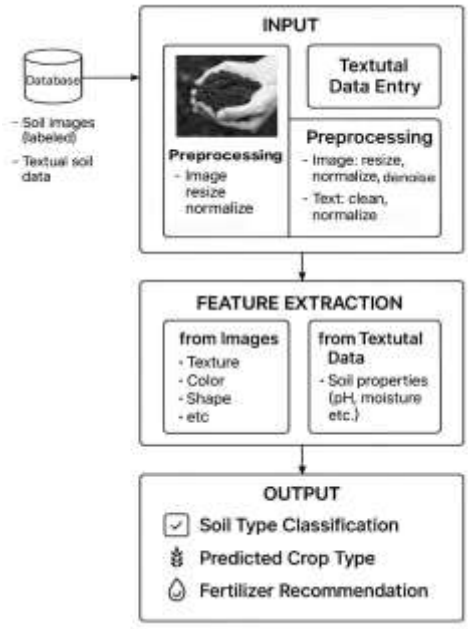

Fig. 2 Implementation

To handle textual soil data—such as pH, moisture, or nutrient content—input values are first cleaned and normalized. This process ensures uniform data formatting and scales the values within an optimal range for model training. Both image and textual data undergo separate pre-processing pipelines but are ultimately fused during the feature extraction stage.

Feature extraction from images involves deep learning techniques to capture essential visual characteristics such as texture, colour, and shape. These features are critical for distinguishing between different soil types visually. Simultaneously, structured features are extracted from textual data, representing chemical and physical soil properties like pH levels, moisture content, and organic matter concentration.

For model training, the extracted features from both modalities are either fed into a joint neural network model or merged using feature fusion techniques. This dual-modality approach enhances the model's ability to learn complex patterns that contribute to soil classification, crop recommendation, and fertilizer suggestion.

The output module of the system delivers three key predictions:

Soil Type Classification – Identifies soil as black, red, alluvial, sandy, etc.

Predicted Crop Type – Suggests the most suitable crops based on soil properties.

Fertilizer Recommendation – Provides optimal fertilizer types and quantities tailored to the crop and soil type.

Unlike traditional image-only classifiers, this hybrid approach ensures higher accuracy and better generalization, especially in agricultural applications where both appearance and physical properties of soil play critical roles. The model's performance is continuously evaluated using a separate validation set, ensuring generalization and adaptability to real-world soil data. Through comprehensive pre-processing, multimodal feature integration, and intelligent output generation, this system offers a robust solution for smart agriculture decision support.

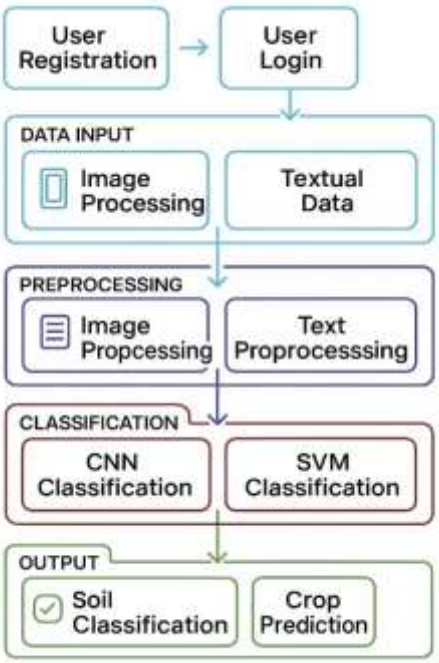

Fig. 3 Implementation Process

V. Results and Discussion

The proposed system for soil classification and crop prediction was developed using a combination of Convolutional Neural Networks (CNN) for analyzing soil images and Support Vector Machine (SVM) for processing soil parameters such as nitrogen, phosphorus, potassium, pH, and moisture levels. A dataset of around 400 labeled samples was used, collected from various agricultural sources and online repositories. The model was trained to classify soil types and recommend suitable crops based on the given inputs.

During testing on the internal dataset, the model showed high accuracy, achieving confidence levels between 85% and 92%. The CNN model effectively captured visual patterns from soil textures, while the SVM handled numerical data with good precision. However, when tested on external data—such as images taken in different lighting conditions or using different devices performance slightly declined, with confidence scores reducing to around 60% to 70%. This variation is likely due to changes in image quality and inconsistent parameter formatting.

To support practical use, a web-based interface was also developed, allowing users to upload soil images and input parameters. The system then predicts the soil type, recommends crops, and displays the prediction confidence, making it easier for farmers to make informed agricultural decisions.

A. Model Performance

The developed system was trained using a combination of soil images and soil parameters such as pH, nitrogen, phosphorus, potassium, and moisture. These two input types helped the model understand both the visual texture of the soil and its chemical properties, making the prediction of soil type and suitable crops more accurate.

A Convolutional Neural Network (CNN) was used to extract meaningful features from the soil images—such as texture, color patterns, and surface structure. In parallel, Support Vector Machine (SVM) was used to classify the soil and predict crops based on both the image features and the soil parameter values. This hybrid approach allowed the model to make more informed decisions by combining visual and numerical data.

During training, the model showed good learning behavior, with accuracy gradually improving over time. It achieved high confidence scores ranging between 80% to 95% on the training dataset, which indicates that the model could understand the relationship between soil conditions and suitable crops effectively.

However, when tested with new data especially soil images taken under different lighting conditions or from unfamiliar regions the model's performance dropped slightly. Confidence scores for predictions sometimes ranged from 50% to 65%, and there were a few incorrect classifications. This shows that the model had difficulty generalizing to completely new inputs, which is common when the training dataset is small or lacks variation.

Despite this, the combined CNN and SVM system provided strong performance on known data and showed promising results on unknown data. It forms a good base for real-world deployment, with potential for improvement as more data becomes available and the model continues to be optimized.

B. Challenges and Future Improvement

To make the soil classification and crop prediction model more reliable and accurate—especially when applied to real-world conditions—several improvements can be considered. One of the most important steps is to increase the size and diversity of the dataset. Collecting more soil data and crop details from different regions, climates, and seasons will allow the model to recognize a broader range of patterns and perform better when it encounters new or unfamiliar inputs.

Enhancing the model architecture can also lead to better performance. For instance, attention mechanisms like Squeeze-and-Excitation (SE) blocks can be added to help the system focus on the most important features in soil images or sensor readings, such as texture differences or moisture content.

Another effective strategy is to combine multiple types of data. Along with images or sensor readings, including parameters like soil nutrients (NPK values), temperature, rainfall, and past crop records can help the system understand the context more deeply and make better predictions. Proper data preprocessing also plays a key role—techniques like background removal, brightness adjustment, and normalization can reduce noise and help the model learn more clearly.

Lastly, maintaining a balanced dataset is essential. Ensuring that all soil types and crop categories are fairly represented prevents the model from becoming biased toward the most common classes. Data augmentation and targeted data collection from underrepresented categories can help address this issue. As more data becomes available over time, the model should also be updated regularly to keep improving its accuracy and adaptability.

VI. Conclusion

Agriculture plays a crucial role in driving the economic development of our country. However, the sector has been slow in adopting modern advancements, particularly in the area of machine learning. It is essential for farmers to become familiar with emerging technologies and innovative practices, as they can significantly boost crop productivity. By integrating machine learning into agricultural processes, we can not only enhance crop yield but also address various challenges faced in farming. These technologies enable accurate predictions and assessments, allowing comparisons across different methods to determine the most effective ones. Ultimately, this leads to improved decision-making and better overall performance in crop production.

VI. References

[1] Y. J. N. Kumar, V. Spandana, V. S. Vaishnavi, K. Neha, and V. G. R. R. Devi, "Supervised machine learning approach for crop yield prediction in agriculture sector," in Proc. Fifth Int. Conf. Communication and Electronics Systems (ICCES 2020), IEEE Conference Record #48766, IEEE Xplore ISBN: 978-1-7281-5371-1, 2020.

[2] P. R. Dhumal, G. R. Sinha, R. B. Pachpor, A. B. Karode, and J. N. Shinde, "Soil classification using machine learning methods," in Proc. 2020 IEEE International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), IEEE Xplore ISBN: 978-1-7281-4865-6, pp. 174–177, 2020.

[3] M. B. Shewale, M. S. Shewale, A. R. Kulkarni, A. Deshmukh, D. Dhade, and D. Gadhave, "Soil classification & crop suggestion using machine learning," in Proc. 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), IEEE Xplore ISBN: 978-1-7281-4167-1, pp. 1–5, 2020.

[4] P. S. Vijayabaskar, R. Sreemathi, and E. Keertanaa, "Crop prediction using predictive analytics," in Proc. Int. Conf. Computation of Power, Energy, Information and Communication (ICCPEIC 2017), IEEE Conference Record #43248, IEEE Xplore ISBN: 978-1-5090-4324-8, 2017.

[5] Z. M. Shaikh, S. S. Gaikwad, and P. R. Tadas, "Crop prediction using machine learning," in Proc. Int. Conf. on Artificial Intelligence and Machine Vision (AIMV 2021), IEEE Conference Record #XXXXXX, IEEE Xplore ISBN: XXXXXXXXXXXX, 2021