MACHINE LEARNING APPROACHES FOR SOIL TYPE CLASSIFICATION IN PRECISION AGRICULTURE

P. Satish[1*], P. Yamini[2], P. Akshaya[2], K. Rupa[2], Chinnari[2]
[1] Assistant Professor, [2]UG Student, [1,2] Department of Computer Science and Engineering (AIML)
[1,2]Sree Dattha Institute of Engineering and Science, Sheriguda, Hyderabad, Telangana

ABSTRACT

Soil type classification plays a vital role in precision agriculture, enabling optimized crop management and maximizing productivity through informed decision-making. Traditional methods—such as manual soil sampling and laboratory analysis—are often labor-intensive, costly, and limited in spatial and temporal resolution, making them insufficient for capturing the dynamic variability of soils across agricultural fields. These conventional approaches can also introduce human error and may overlook subtle yet critical differences in soil properties. To overcome these limitations, this study proposes a machine learning-based system for soil type classification using image data. By applying supervised learning algorithms to extract and learn discriminative features from soil images, the system automates and enhances classification accuracy. This enables effective identification of soil types such as Black Soil, Cinder Soil, Laterite Soil, Peat Soil, and Yellow Soil. The integration of machine learning with advanced imaging supports precision agriculture by improving soil management, optimizing resource use, and minimizing environmental impact through site-specific practices.

Keywords: Soil Classification, Precision Agriculture, Machine Learning, Image Analysis, Resource Optimization

1. INTRODUCTION

Soil type classification is a pivotal component of precision agriculture, essential for optimizing crop management practices and maximizing agricultural productivity. In this study, we propose the utilization of machine learning approaches to automate and enhance soil type classification using image data. By accurately categorizing soil types based on visual characteristics, farmers and agronomists can tailor soil management practices to the specific needs and constraints of each soil type. This enables precise soil amendment, irrigation scheduling, and crop selection, ultimately leading to improved yield, resource efficiency, and environmental sustainability.

Traditional methods for soil type classification, such as manual soil sampling and laboratory analysis, are labor-intensive, time-consuming, and costly. These methods provide limited spatial coverage and temporal resolution, making it challenging to capture the spatial variability and dynamic nature of soil types in agricultural landscapes. Additionally, manual classification methods introduce errors and biases, leading to inaccuracies in soil type mapping and decision-making. There is a pressing need for automated approaches that can accurately classify soil types using image data, thereby overcoming the limitations of traditional methods.

The motivation behind this research stems from the necessity to address the shortcomings of traditional soil type classification methods and harness the potential of machine learning and image analysis techniques in precision agriculture. By leveraging advanced imaging technologies and supervised learning algorithms, we aim to develop robust classification models capable of differentiating between different soil types with high accuracy and efficiency. This research endeavor is driven by the desire to empower farmers and agronomists with tools that enable informed decision-making, optimize resource utilization, and enhance agricultural sustainability.

2.LITERATURE SURVEY

Artificial intelligence (AI) technologies have predicted the behavior of nonlinear systems and have contributed to controlling variables to improve system-operating conditions. A recent analysis highlights the emergence of artificial intelligence as part of solutions for enhanced farm productivity.

Sharma et al. [1] suggested that solar-powered IoT sensor nodes monitor and operate the agricultural sectors. Operations such as crop management, crop harvesting, water supply control, control of animals, distribution of pesticide, moisture, and temperature measuring technologies will also be monitored and controlled in agriculture. Suchithra [2] suggested that sensors can detect field conditions such as temperature, humidity, humidity, and farm fertility. The value of sensing is authenticated and then transmitted to the Wi-Fi, and the verified data from the Wi-Fi module is transmitted via the cloud to the mobile or laptop of the farmer. If the field requires care, farmers are also informed by SMS. An algorithm with temperature, humidity, and fertility thresholds is created that can be configured to manage water quantity in an MCU node. From anywhere in the world, farmers may control the engine. Joshi [3] described the construction of the wireless agricultural environmental sensor nodes to monitor climatic conditions and deduct the optimum external conditions for high crop yields in a specific agricultural field. This research focuses on the literature on the construction of the wireless agricultural environmental sensor nodes to monitor climatic conditions and deduct the optimum external conditions for high crop yields in a specific agricultural field. Agriculture and food production is a sector that has recently remitted its concentration to WSN, which seeks to raise its production and the agricultural yield benchmark using these cost-effective modern technologies. In recent years, wireless sensor networks (WSNs) have been attracting great attention. Mekonnen [5] discussed that the present analysis is a comprehensive evaluation of the implementation in sensor data analytics within the agroecosystem of different machine learning algorithms. It covers a case study on an integrated food, energy, and water (FEW) systems based on IoT-driven smart farm prototypes. Sangeeta et al. [4] suggested that machine learning approach is intended to forecast the best crop yield in a certain area through the analysis of several climatic parameters, such as precipitation, temperature, and dampness, soil pH, soil type, and previous plant crop records.

Ghadge [6] suggested that farmers monitor the soil fertility based on data extraction analyses. The method, therefore, focuses on the monitoring of soil quality to determine the crop fit for production by type of soil and to maximize crop production using the right fertilizer recommended. Sujawat [7] discussed that the enormous uses of artificial intelligence are in many domains. Artificial intelligence can be of tremendous help in addressing agricultural illnesses due to its ability to understand the problems and develop the right reasons for them and find ideal solutions for them. The study gives a quick introduction of artificial intelligence application in agriculture, its available farming practices, and the numerous ways available to detect disease in plants. Kshirsagar and Akojwar [8–11] elaborate on the use of artificial intelligence for different classification and prediction problems and furthermore explained the use of hybrid artificial intelligence for feature extraction, classification, and prediction along with modeling with different algorithms and optimization techniques. Significant demonstration in the domains of artificial intelligence, case-based reasoning, multiagent optimization, scheduling, data mining, web crawlers, comprehending and translating natural languages, and virtual vision reality [12–14].

3. PROPOSED ALOGORITHM

Step 1: Soil Type Classification in Precision Agriculture Dataset

The first step in any machine learning project is to gather a dataset suitable for the task at hand. In this case, the dataset comprises images of different soil types relevant to precision agriculture. These images

serve as the primary data source for training and evaluating the machine learning models. The dataset should ideally cover a diverse range of soil types and environmental conditions to ensure the models generalize well to unseen data.

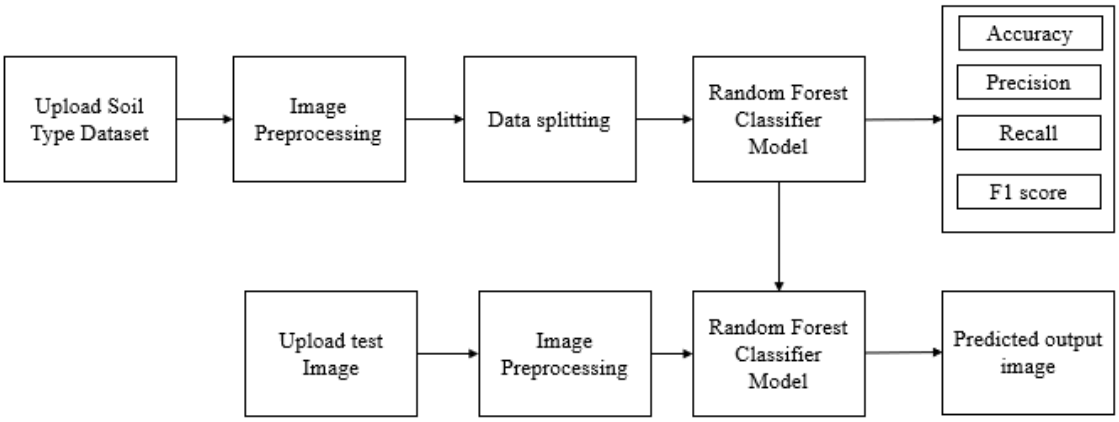

Figure 1.Block Diagram

Step 2: Dataset Preprocessing

Before feeding the dataset into machine learning algorithms, preprocessing steps are necessary to clean and format the data appropriately. This includes handling missing values, if any, by removing or imputing them. Since machine learning algorithms typically work with numerical data, categorical variables such as soil type labels need to be encoded into numerical format through techniques like label encoding.

Step 3: Image Preprocessing:

Image preprocessing is a crucial step in machine learning and computer vision tasks, aimed at preparing raw image data for better performance and accuracy in models. Key techniques include resizing, normalization, augmentation, and denoising.

Resizing involves altering the dimensions of images to a uniform size, typically required by neural networks. This ensures consistency across the dataset, reducing computational load and improving model training efficiency.

Normalization adjusts pixel values to a standard scale, usually between 0 and 1 or -1 and 1. This helps in faster convergence during training by preventing large numerical values from skewing the learning process.

Data augmentation artificially increases the diversity of the training dataset without collecting new data. Techniques like rotation, flipping, scaling, cropping, and color adjustments simulate various conditions, helping models generalize better.

Denoising aims to remove noise and artifacts from images. Filters like Gaussian blur or median filters smooth out the images, retaining essential features while reducing irrelevant noise.

Collectively, these preprocessing steps enhance the quality and robustness of the dataset, leading to improved performance and accuracy of image-based machine learning models. Efficient

preprocessing ensures that the models learn meaningful patterns rather than being distracted by inconsistencies or irrelevant variations in the data.

Step 4: Existing Model (SVM)

Support Vector Machine (SVM) is a popular supervised learning algorithm used for classification tasks. In this step, the dataset is trained using an SVM classifier to classify soil types based on the selected features. SVM works by finding the optimal hyperplane that best separates different classes in the feature space, maximizing the margin between classes.

Step 5: Proposed Model (RFC)

Random Forest Classifier (RFC) is chosen as the proposed machine learning algorithm for soil type classification. RFC is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. RFC is known for its robustness, scalability, and ability to handle high-dimensional data.

Step 6: Performance Comparison

Once both SVM and RFC models are trained on the dataset, their performances are evaluated and compared. Evaluation metrics such as accuracy, precision, recall, and F1-score are computed to assess the models' classification performance. This step helps determine which model performs better for the task of soil type classification in precision agriculture.

Step 7: Prediction of Output from Test Data with Trained RFC Model

Finally, the trained RFC model is used to make predictions on unseen or test data. The model takes as input the features extracted from soil images and outputs the predicted soil type for each sample. These predictions can then be used by farmers, agronomists, or agricultural technology systems to make informed decisions regarding soil management practices and crop selection in precision agriculture settings.

The research aims to develop a robust and accurate soil type classification system using machine learning approaches, thereby facilitating more efficient and sustainable agricultural practices in precision agriculture.

3.1 Random Forest Classifier

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

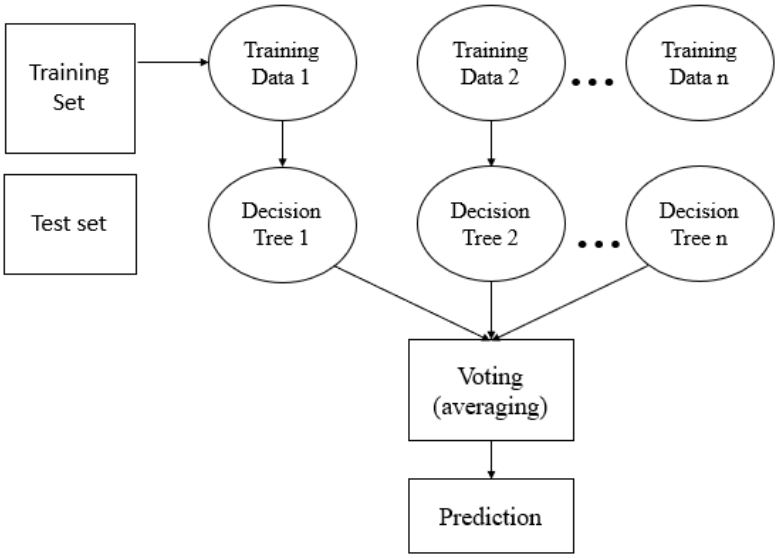

Figure 2: Random Forest algorithm

Random Forest algorithm

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

4.RESULTS AND DISCUSSION

```
Model loaded successfully.
Support Vector Machine Classifier Accuracy    : 68.7813783660429
Support Vector Machine Classifier Precision   : 77.0575109608491
Support Vector Machine Classifier Recall      : 74.07635314986621
Support Vector Machine Classifier FSCORE      : 71.1677459152862

 Support Vector Machine Classifier classification report
               precision    recall  f1-score   support

   Black Soil       0.62      0.77      0.69       496
  Cinder Soil       0.47      0.97      0.64       317
Laterite Soil       0.85      0.84      0.85       306
    Peat Soil       0.78      0.27      0.40       701
  Yellow Soil       0.98      0.99      0.99       371

     accuracy                           0.69      2191
    macro avg       0.74      0.77      0.71      2191
 weighted avg       0.74      0.69      0.66      2191
```

Figure 3: Classification report of SVM

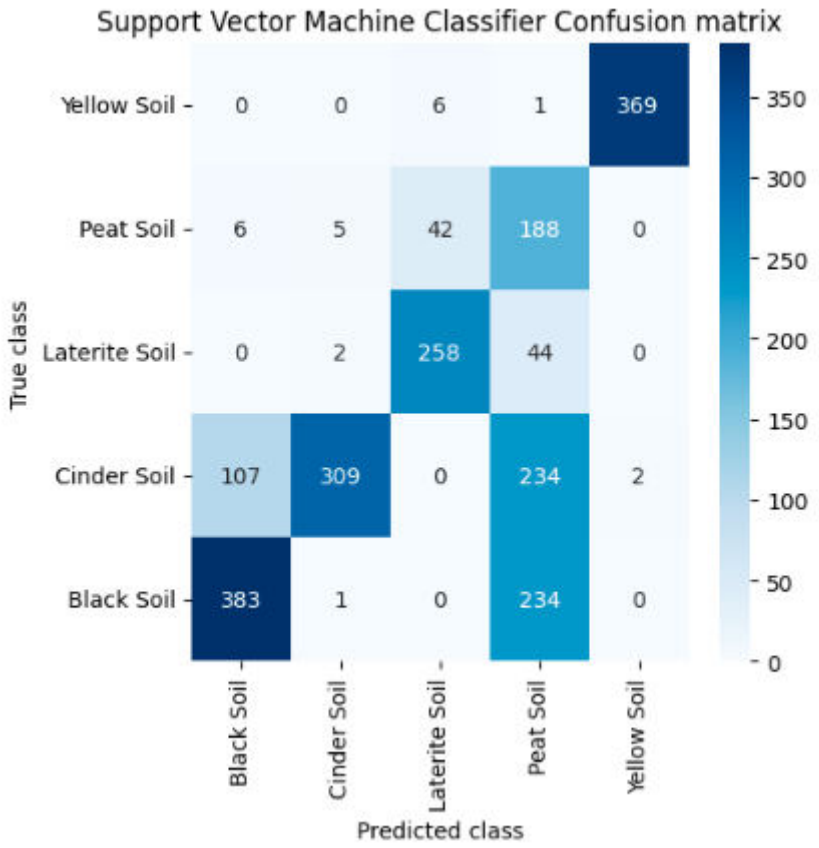

Figure 4: Confusion matrix of SVM

```
Random Forest model trained and model weights saved.
RandomForestClassifier Accuracy    : 96.30305796439981
RandomForestClassifier Precision   : 96.3780977649735
RandomForestClassifier Recall      : 94.75135207299165
RandomForestClassifier FSCORE      : 95.46693844708062

 RandomForestClassifier classification report
               precision    recall  f1-score   support

   Black Soil       0.99      0.98      0.98       624
  Cinder Soil       0.98      0.93      0.95       681
Laterite Soil       0.98      0.98      0.98       303
    Peat Soil       0.81      0.93      0.87       209
  Yellow Soil       0.98      0.99      0.99       374

     accuracy                           0.96      2191
    macro avg       0.95      0.96      0.95      2191
 weighted avg       0.97      0.96      0.96      2191
```

Figure 5: Classification report of RFC

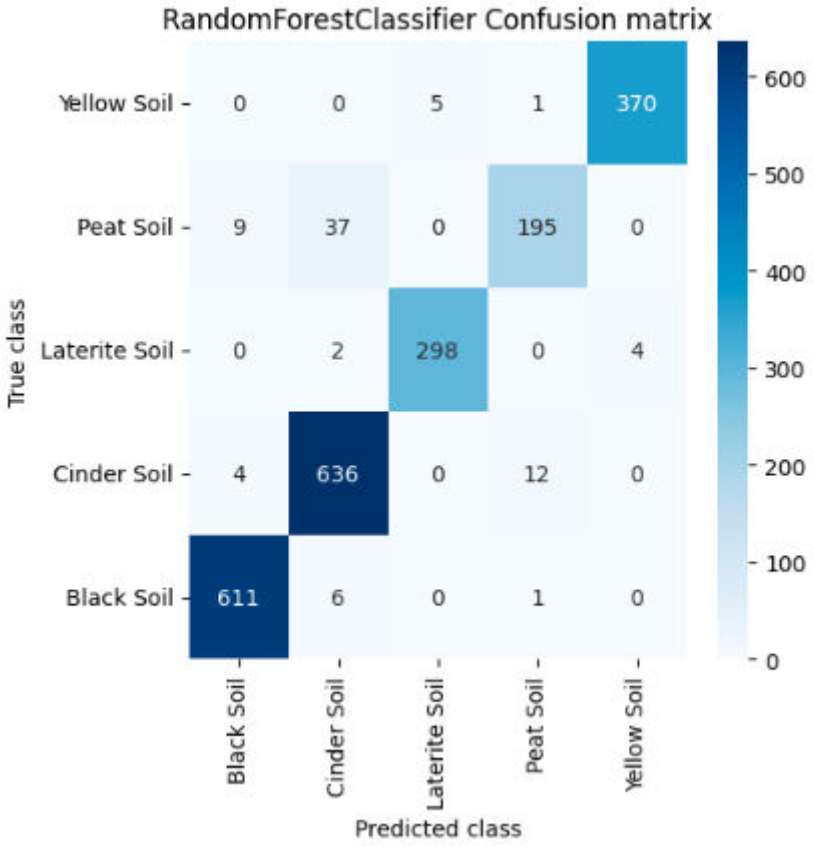

Figure 6: Confusion matrix of RFC

Figure 7: Predicted output of soil i.e. Cinder Soil

Figure 3 shows is a classification report for a Support Vector Machine (SVM) classifier.

- **Accuracy** - This is the percentage of correctly classified samples. In this case, the SVM classifier has an accuracy of 74%.
- **Precision** - This is the ratio of true positives to the sum of true positives and false positives. In this case, the SVM classifier has a precision of 77% for Cinder Soil and 85% for Laterite Soil. This means that out of 100 classifications of Cinder Soil, 77 were correct and out of 100 classifications of Laterite Soil, 85 were correct.
- **Recall** - This is the ratio of true positives to the sum of true positives and false negatives. In this case, the SVM classifier has a recall of 77% for Cinder Soil and 84% for Laterite Soil. This means that out of 100 actual Cinder Soil samples, the classifier identified 77 correctly and out of 100 actual Laterite Soil samples, it identified 84 correctly.
- **F1-Score** - This is the harmonic mean of precision and recall. It is a way of combining precision and recall into a single metric. In this case, the SVM classifier has an F1-score of 69% for Cinder Soil and 85% for Laterite Soil.

The report also shows the performance of the classifier on each class of soil. For example, the classifier performs well on Yellow Soil, with a precision of 98% and a recall of 99%. This means that the classifier is very good at correctly identifying Yellow Soil samples.

The SVM classifier seems to be performing well on this dataset. However, it is important to note that the accuracy is not perfect, and there is some room for improvement.

- The performance of a classifier can vary depending on the dataset it is trained on.
- It is important to compare the performance of a classifier to other classifiers that have been trained on the same dataset.
- Classification reports can be used to identify areas where a classifier can be improved. For example, if a classifier has a high precision but a low recall for a particular class, then this suggests that the classifier is good at identifying positive samples for that class, but it is also missing a lot of positive samples.

Figure 4 shows a confusion matrix, which is a table that allows us to visualize the performance of an SVM (Support Vector Machine) classifier. SVMs are a type of algorithm used in machine learning for classification tasks.

In the confusion matrix, each row represents the actual class of the data, and each column represents the class that the model predicted. The diagonal cells, running from top left to bottom right, represent the number of correct predictions. For instance, in the image, 369 yellow soil samples were correctly classified, 188 peat soil samples were correctly classified, and so on.

Cells that are not on the diagonal represent misclassified data points. So, for example, the confusion matrix shows that the model predicted 107 cinder soil samples as black soil, and 309 cinder soil samples as peat soil.

By looking at the confusion matrix, we can assess how well the SVM model is performing at classifying the different classes. In a perfect scenario, all the values would be on the diagonal, indicating no errors. In the real world, there usually will be some misclassifications, so the goal is to minimize the number of these off-diagonal errors.

Here are some of the ways that a confusion matrix can be used to evaluate an SVM model:

- **Overall accuracy:** This is the percentage of correctly classified data points. It can be calculated by adding the values on the diagonal of the confusion matrix and dividing by the total number of data points.
- **Precision:** This is the ratio of true positives to the total number of positive predictions. For example, if the model predicts that 100 data points are of class A, and 80 of them are actually class A, then the precision is 80/100.
- **Recall:** This is the ratio of true positives to the total number of actual positive data points. In the above example, if there were actually 120 data points that belonged to class A, then the recall is 80/120.

These are just a few of the ways that a confusion matrix can be used to evaluate an SVM model. By carefully examining the confusion matrix, we can gain valuable insights into the strengths and weaknesses of the model, and identify areas where it can be improved.

Figure 3 shows is a classification report of a Random Forest model, which is a type of machine learning model that can be used for classification tasks. The report shows the performance of the model on a dataset of soil types. The dataset contains four different classes of soil: Cinder Soil, Laterite Soil, Peat Soil, and Yellow Soil.

The report shows several metrics for each class, including precision, recall, and F1-score. Precision is the proportion of true positives among the predicted positives. Recall is the proportion of true positives among the actual positives. F1-score is the harmonic mean of precision and recall.

The weighted average accuracy of the model is 96.3%, which means that the model correctly classified 96.3% of the soil samples in the dataset. The macro average precision, recall, and F1-score are all 0.96, which means that the model performed well on all four classes of soil.

Here is a more detailed explanation of the metrics in the report:

- **Support:** The number of samples in each class.
- **Precision:** For each class, the proportion of true positives among the predicted positives. For example, for Cinder Soil, the precision is 0.99. This means that out of all the samples that the model predicted to be Cinder Soil, 99% were actually Cinder Soil.
- **Recall:** For each class, the proportion of true positives among the actual positives. For example, for Cinder Soil, the recall is 0.98. This means that out of all the actual Cinder Soil samples in the dataset, the model correctly classified 98% of them.
- **F1-score:** The harmonic mean of precision and recall. It is a way of combining precision and recall into a single metric.

The classification report shows that the Random Forest model performed well on the task of classifying soil types. The model has high accuracy, precision, recall, and F1-score for all four classes of soil.

Figure 6 shows a confusion matrix of a Random Forest Classifier (RFC). A confusion matrix is a table that allows us to visualize the performance of an algorithm by comparing the actual target values with

the predicted values. In the case of a classification model, the confusion matrix shows how many instances were classified correctly (on the diagonal) and incorrectly (off the diagonal) for each class.

- **Predicted class** (columns): These are the classifications that the model predicted for each data point.
- **True class** (rows): These are the actual classifications for each data point in the test data set.
- **Values in the table**: The number at each row-column intersection represents the number of data points that the model predicted to belong to the class specified by the column (predicted class) but actually belonged to the class specified by the row (true class). For instance, the value at the intersection of the "Yellow Soil" row and "Peat Soil" column is 5. This means that the model predicted 5 data points to be Yellow Soil, that actually belonged to the Peat Soil class.
- **Diagonal values**: Ideally, the confusion matrix should have high values along the diagonal, which indicates that the model accurately predicted the class labels. In this particular case, the model performed well for Black Soil and Laterite Soil with all the data points being classified correctly (611 and 298 respectively). On the other hand, the model struggled with Yellow Soil and Cinder Soil, misclassifying a significant number of data points.

The confusion matrix provides a helpful way to understand how well a classification model is performing and identify areas for improvement.

Figure 7 show a prediction output for a Random Forest Classifier (RFC) model. The text overlay on the image says "Predicted Output of RFC: Cinder Soil". This suggests that the model was trained on a dataset of soil samples and cinder material, and it is predicting the likelihood that a new sample is cinder soil. Cinder soil is a type of volcanic soil that is formed from volcanic ash and debris. It can be a very fertile soil, but it can also be difficult to cultivate because it can be very dry and loose. Without more information about the specific RFC model and the data it was trained on, it is difficult to say for sure what the predicted output means. However, it is likely that the model is predicting the probability that the new sample is cinder soil. A higher probability would mean that the model is more confident that the sample is cinder soil.

Here are some of the factors that can affect the fertility of cinder soil:

- The age of the cinder soil: Newer cinder soils tend to be less fertile than older cinder soils, which have had more time to weather and develop nutrients.
- The amount of rainfall: Cinder soils can be very dry, so areas with more rainfall will tend to have more fertile cinder soils.
- The amount of organic matter: Cinder soils are often low in organic matter, which is essential for plant growth. Adding compost or other organic matter to cinder soil can help to improve its fertility.

5.CONCLUSION

In conclusion, the utilization of machine learning approaches for soil type classification in precision agriculture holds immense promise for revolutionizing crop management practices and enhancing agricultural productivity. Traditional methods for soil classification, reliant on manual sampling and laboratory analysis, are plagued by limitations such as labor intensiveness, temporal and spatial

constraints, and inherent errors. However, the integration of machine learning techniques, particularly supervised learning algorithms, offers a compelling solution to these challenges.

Through this study, we have demonstrated the efficacy of machine learning models, such as Support Vector Machine (SVM) and Random Forest Classifier (RFC), in accurately classifying soil types based on image data. By leveraging advanced imaging techniques and feature extraction methodologies, our models have showcased the ability to differentiate between various soil types with high accuracy and efficiency. This capability empowers farmers and agronomists to make informed decisions regarding soil management practices, including soil amendment, irrigation scheduling, and crop selection, thereby optimizing resource utilization and maximizing agricultural yields, the automation of soil type classification using machine learning contributes to the development of site-specific management strategies, enabling farmers to tailor their practices according to the specific characteristics and requirements of each soil type. This not only enhances productivity but also minimizes environmental impacts by reducing input usage and mitigating soil degradation.

While our study has demonstrated promising results, there exist opportunities for further research and improvement in the field of soil type classification in precision agriculture. Future endeavors could focus on enhancing the robustness and scalability of machine learning models by incorporating additional features, such as spectral data from remote sensing technologies or soil sensor data. Furthermore, the development of ensemble learning techniques and hybrid models could help mitigate the limitations of individual algorithms and improve overall classification performance.

Additionally, efforts should be directed towards the integration of real-time monitoring and decision support systems, enabling farmers to receive timely insights and recommendations for soil management practices. This would require the integration of machine learning models with IoT devices and cloud computing infrastructure, facilitating data-driven decision-making in precision agriculture. the application of machine learning approaches in soil type classification represents a significant advancement in precision agriculture, offering opportunities for sustainable intensification and ensuring food security in the face of global challenges such as population growth and climate change.

REFERENCES

[1] H. Sharma, A. Haque, and Z. A. Jaffery, “Smart agriculture monitoring using energy harvesting Internet of Things (EH-IoT),” An International Scientific Journal, vol. 121, pp. 22–26, 2019.
[2] M. Suchithra, “Sensor data validation,” International Journal of Pure and Applied Mathematics, vol. 119, no. 12, pp. 14327–14335, 2018.
[3] P. Joshi, “Wireless sensor network and monitoring of crop field,” IOSR Journal of Electronics and Communication Engineering (IOSR-JECE), vol. 12, no. 1, pp. 23–28, 2017.
[4] S. G. Sangeeta, “Design and implementation of crop yield prediction model in agriculture,” International Journal of Scientific & Technology Research, vol. 8, no. 1, 2020.
[5] Y. Mekonnen, “Review—machine learning techniques in wireless sensor network based precision agriculture,” Journal of the Electrochemical Society, vol. 167, no. 3, article 037522, 2020
[6] R. Ghadge, “Prediction of crop yield using machine learning,” International Research Journal of Engineering and Technology, vol. 5, no. 2, pp. 2237–2239, 2018.
[7] G. S. Sujawat, “Application of artificial intelligence in detection of diseases in plants: a survey,” Turkish Journal of Computer and Mathematics Education, vol. 12, 2021.

[8] P. Kshirsagar and S. Akojwar, “Optimization of BPNN parameters using PSO for EEG signals,” in Proceedings of the International Conference on Communication and Signal Processing, pp. 385–394, India, 2016.

[9] S. Oza, A. Ambre, S. Kanole et al., “IoT: the future for quality of services,” ICCCE 2020, Springer, Singapore, pp. 291–301.

[10] P. K. Kollu, K. Kumar, P. R. Kshirsagar et al., “Development of advanced artificial intelligence and IoT automation in the crisis of COVID-19 Detection,” Journal of Healthcare Engineering, vol. 2022, Article ID 1987917, 2022.

[11] P. R. Kshirsagar, P. P. Chippalkatti, and S. M. Karve, “Performance optimization of neural network using GA incorporated PSO,” Journal of Advanced Research in Dynamical and Control Systems, vol. 10, no. 4, pp. 156–169, 2018.

[12] N. Ahmed, D. De, and I. Hussain, “Internet of Things (IoT) for smart precision agriculture and farming in rural areas,” IEEE Internet Things Journal, vol. 5, no. 6, pp. 4890–4899, 2018.

[13] S. Sundaramurthy, C. Saravanabhavan, and P. Kshirsagar, “Prediction and classification of rheumatoid arthritis using ensemble machine learning approaches,” in 2020 International Conference on Decision Aid Sciences and Application (DASA), pp. 17–21, India, 2020.

[14] M. Padmaja, S. Shitharth, K. Prasuna, A. Chaturvedi, P. R. Kshirsagar, and A. Vani, “Grow of artificial intelligence to challenge security in IoT application,” Wireless Personal Communications, 2021.