Crop Yield Prediction Using Machine Learning

Divya Dharshini S[1], Preetha R[2], Dr. S. Shahar Banu[3]
[1, 2] Dept of Computer Applications
[2]Associate Professor, Dept of Computer Applications
[1, 2] B.S. Abdur Rahman Crescent Institute of Science and Technology, India

Abstract- *In this paper the modern agriculture, optimizing plant growth and productivity relies heavily on understanding the complex interplay between soil conditions and plant requirements study uses machine learning algorithms to predict optimal plant species for specific soil conditions based on soil properties like water content, nutrients, pH, and organic matter. Supervised techniques like decision trees, random forests, and SVMs are employed on a dataset of soil characteristics and plant growth outcomes. Model performance is rigorously evaluated. The goal is a decision-support tool for crop selection to enhance agricultural productivity while accounting for soil variation.*

Keywords- decision trees, random forests, support vector machines, Machine learning, CNN.

I. INTRODUCTION

In modern agriculture, the quest for enhanced plant growth and productivity necessitates a deep understanding of the intricate interplay between soil conditions and plant requirements.Soil serves as the foundation for crop cultivation, providing essential nutrients and water while also serving as a habitat for root development. However, the heterogeneity of soil properties across different regions poses a significant challenge for farmers and agricultural experts in determining the most suitable plant species for specific soil conditions. To address this challenge, this study proposes a pioneering approach that leverages machine learning algorithms to assess plant suitability in diverse soil conditions, considering crucial factors such as soil water content and nutrient levels. By amalgamating data from various sources, including soil composition, moisture content, and nutrient profiles, the study endeavour to develop predictive models capable of discerning which plant species thrive optimally in specific soil profiles. The methodology employs supervised learning techniques, encompassing decision trees, random forests, and support vector machines, to forecast the suitability of different plant species under varying soil conditions. Key features such as soil pH, moisture content, organic matter, and nutrient composition are extracted and utilized as input variables for the machine learning models. The models are trained on a comprehensive dataset comprising soil characteristics and corresponding plant growth outcomes, thereby facilitating robust predictive capabilities.The efficacy of each algorithm is rigorously evaluated through extensive experimentation and cross-validation, scrutinizing performance metrics including accuracy, precision, recall, and F1-score.Furthermore, feature importance analysis is conducted to delineate the pivotal factors influencing plant suitability across diverse soil conditions.In contemporary agriculture, selecting crops suited to specific soil conditions is crucial for boosting yield and promoting sustainability. Farmers must match crops with soil compositions to optimize productivity and reduce resource wastage. This study utilizes convolutional neural network (CNN) algorithms to analyze plant suitability based on soil attributes like water and nutrient levels. By integrating these factors, we aim to provide tailored insights for farmers. We collect extensive soil data from diverse regions and employ CNN techniques for robust analysis. Through rigorous experimentation, we evaluate the precision and reliability of our methodology in predicting crop suitability. Ultimately, this research aims to empower farmers with actionable insights for enhanced agricultural productivity and sustainability, fostering a more environmentally conscious approach to farming practices.

II. LITERATURE SURVEY

Crop yield prediction is a multifaceted domain within agricultural research that has garnered increasing attention due to its significance in optimizing agricultural practices, ensuring food security, and mitigating risks associated with crop failures. Machine learning techniques have emerged as pivotal tools in this field, offering the potential to leverage vast datasets and complex relationships inherent in agricultural systems for accurate yield forecasting.A seminal study by Lu et al. (2020) provides a comprehensive review and meta-analysis specifically focusing on the application of deep learning techniques in crop yield prediction. The study evaluates the effectiveness of various deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), in capturing spatiotemporal patterns from diverse data sources including remote sensing imagery, weather data, and soil properties. The meta-analysis conducted by Lu et al. highlights the superior performance of deep learning models compared to traditional machine learning approaches in predicting crop yields,

emphasizing their ability to handle complex nonlinear relationships and high-dimensional input data.Expanding on the broader landscape of machine learning applications in precision agriculture, Mouazen et al. (2019) provide a comprehensive review covering various machine learning algorithms employed for crop yield prediction and nitrogen status estimation. The review encompasses methodologies ranging from classical regression techniques to more advanced ensemble methods and neural networks. Mouazen et al. emphasize the importance of integrating multi-source data, including spectral reflectance data, geographic information system (GIS) data, and machine sensor data, to develop robust predictive models capable of accurately estimating crop yields and assessing nutrient status in agricultural fields.Singh et al. (2019) contribute further insights into the methodological aspects of crop yield prediction by exploring a wide array of machine learning and data mining approaches. The review encompasses techniques such as feature selection, dimensionality reduction, and model evaluation, highlighting their importance in enhancing the performance and interpretability of predictive models. Singh et al. stress the need for rigorous data pre-processing techniques to handle missing values, outliers, and noise inherent in agricultural datasets, as well as the significance of domain knowledge integration to improve model generalization and transferability across different agricultural settings.Building upon these foundational studies, Madhura et al. (2021) conduct a meticulous literature review categorizing existing research based on crop types, machine learning algorithms, and data sources utilized for crop yield prediction. The review identifies common challenges faced by researchers in the field, including data scarcity, model interpretability, and scalability issues. Madhura et al. propose potential solutions to address these challenges, such as the integration of remote sensing data, crowdsourced data, and advanced data fusion techniques to augment traditional agricultural datasets.Furthermore, Raza et al. (2019) emphasize the interdisciplinary nature of crop yield prediction, advocating for collaborative efforts between agronomists, computer scientists, and data scientists.

Their study underscores the importance of integrating diverse data sources, including environmental factors, genomic information, and socioeconomic variables, to develop comprehensive predictive models capable of capturing the complex interactions between agronomic practices, environmental conditions, and crop physiology.

In conclusion, the collective body of work reviewed herein elucidates the current state-of-the-art methodologies, challenges, and future directions in crop yield prediction using machine learning. These studies provide valuable insights for researchers and practitioners seeking to enhance the accuracy and reliability of crop yield forecasts, thereby facilitating informed decision-making in agriculture and contributing to global food security efforts.

III. PROPOSED SYSTEM

This paper explores Convolutional Neural Networks (CNNs) to teach and identify objects in pictures. We start by gathering images and annotating them.Using Python, we write code and applying techniques to improve recognition accuracy.We will compare different algorithms to identify the best approaches, using Python for analysis.

Documentation is a key part of our process, and we'll share reports and visualizations to make our findings accessible.Collaboration within our team is crucial for problem-solving and idea generation. By the end of the project, we aim to understand CNNs better and explore applications.

By analysis the soil water content, nutrients content we can decide the growth of the plant spices in the soil. Finding out whether the crop can grow well or not in the specified soil. Key features are including soil, pH, moisture content, organic matter, and nutrients compositions are extracted and used as the input variables. The main aim is soil testing and water content testing.

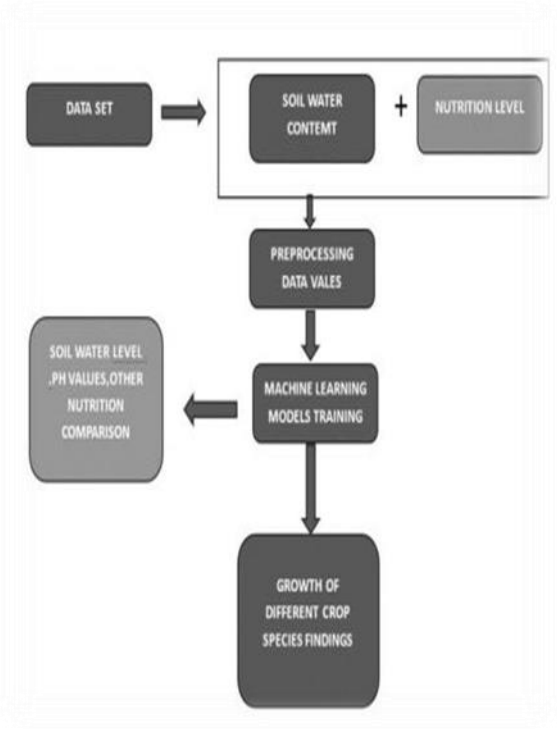

Fig: 3.1. Architecture diagram of crop yield prediction.

The figure 3.1 refers the architecture diagram of crop yield prediction that contains the dataset of soil water contents and nutrients.

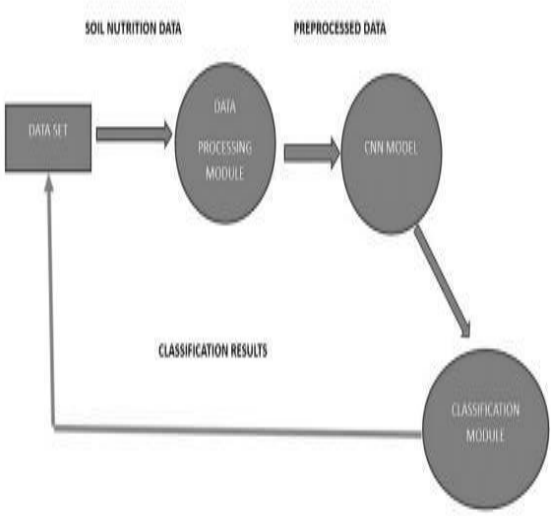

Fig: 3.2 Data flow diagram of crop yield prediction.

The figure 3.2 refers the dataflow diagram of the crop yield prediction that contain the data collection and data pre-processing.

The dataset focuses on collecting soil condition data from sources like agricultural research databases, soil science publications, and experimental farms. The data includes soil properties like pH, moisture content, nutrient levels, and plant growth outcomes for different crops.

Ensuring data integrity and quality assurance is crucial for accurate analysis.This cleans and prepares data for analysis. It includes features.

Features selection techniques like correlation analysis or dimensionality reduction methods like PCA may be used to reduce computation and improve model performance.We collect extensive soil data from diverse regions and employ CNN techniques for robust analysis. Through rigorous experimentation, we evaluate the precision and reliability of our methodology in predicting crop suitability.

	N	P	K	temperature	humidity	ph	rainfall
0	90	42	43	20.879744	82.002744	6.502985	202.935536
1	85	58	41	21.770462	80.319644	7.038096	226.655537
2	60	55	44	23.004459	82.320763	7.840207	263.964248
3	74	35	40	26.491096	80.158363	6.980401	242.864034
4	78	42	42	20.130175	81.604873	7.628473	262.717340

Fig: 3.3. Dataset of crop yield prediction.

The figure 3.3 refers the dataset that contains the corresponding values.

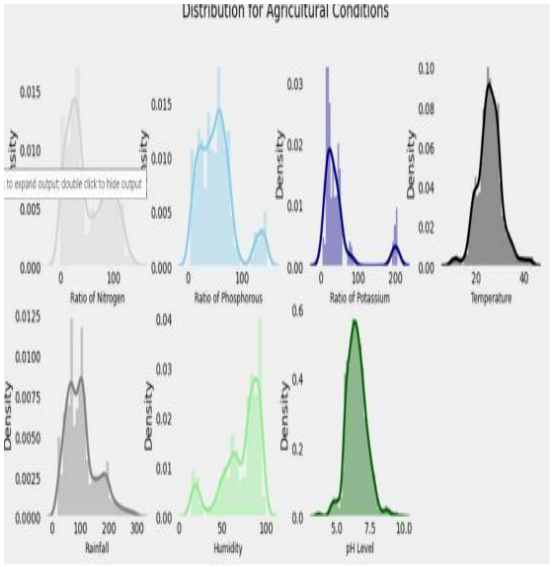

Fig: 3.4. Graph of density and rainfall.

From figure 3.4 the Subplots to visualize the distribution of different agricultural conditions using seaborndistplot. It shows the frequency distribution for various data attributes: nitrogen, phosphorous, potassium, temperature, rainfall, humidity, and pH level. Each subplot uses a different color for differentiation, and the overall plot has a title indicating its focus on agricultural conditions.

Random Forest:

Random Forest is a machine learning method that uses multiple decision trees to make predictions, combining their outputs to increase accuracy and reduce overfitting. In crop yield prediction, it takes various factors like soil composition, climate, rainfall, and temperature to estimate crop yields. Its robustness, accuracy, and ability to identify

important features makeit a popular choice for agricultural applications.

Steps Involved in Random Forest Algorithm:

Step 1: Import the necessary libraries

```python
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
```

Step 2: Prepare your data

```python
# Assuming you have your data stored in X (features) and y (target variable)
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Step 3: Create a Random Forest classifier

```python
# Initialize the Random Forest classifier
rf_classifier = RandomForestClassifier(n_estimators=100, random_state=42)
```

Step 4: Train the classifier on the training data

```python
# Fit the classifier to the training data
rf_classifier.fit(X_train, y_train)
```

Step 5: Make predictions on the testing data

```python
# Use the trained classifier to make predictions
y_pred = rf_classifier.predict(X_test)
```

Step 6: Evaluate the performance of the model

```python
# Calculate the accuracy of the model
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

CNN:

Convolutional Neural Network(CNNs) are used in crop yield prediction by analysing spatial data like satellite images to identify patterns related to crop health, growth stages, and other factors.

This allows CNNs to estimate yields and provide insights into crop conditions. Their ability to process complex spatial data makes them useful for real-time monitoring and prediction in agriculture.

IV. IMPLEMENTATION

The snippet creates a bar plot comparing the accuracy of two machine learning models, Random Forest and CNN, using the Seaborn library.

It uses the specified accuracies (y) and adds a title and y-axis label to visualize the accuracy comparison between the two models in a straightforward manner.

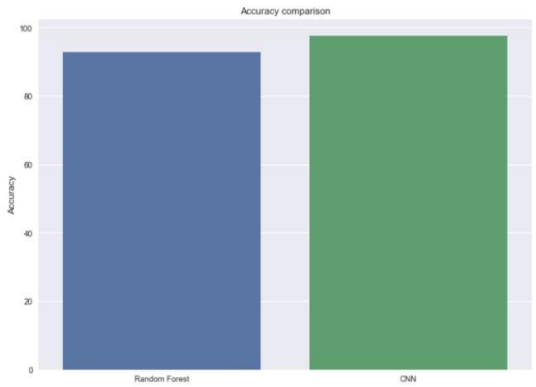

Fig: 4.1. Accuracy comparison.

The figure 4.1 refers the accuracy of random forest and CNN.

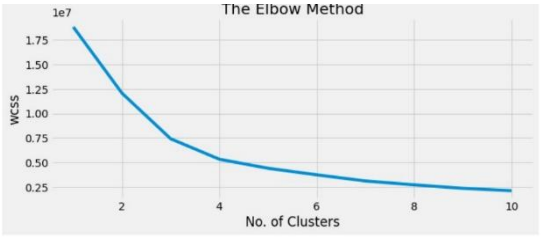

Fig: 4.2. The graph of crop yield prediction.

The figure 4.2 refers the number of clusters.

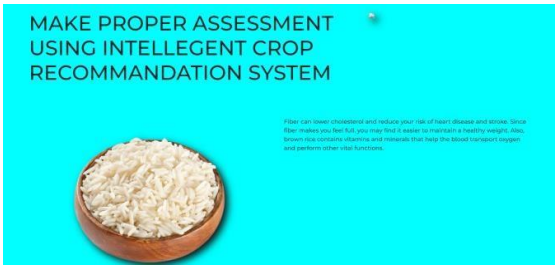

Fig: 4.3. Implementing the crop yield prediction system.

The figure 4.3 refers the implement of prediction system on an application using tools like flask run to build the site.

V. RESULT

The results provided insights into key factors affecting plant growth across various soil profiles, helping to guide agricultural practices. This research promotes sustainable and efficient crop cultivation through informed decision-making.

Fig: 5.1. The result of the crop yield prediction using machine learning.

The figure 5.1 refers the final outcome of the project.

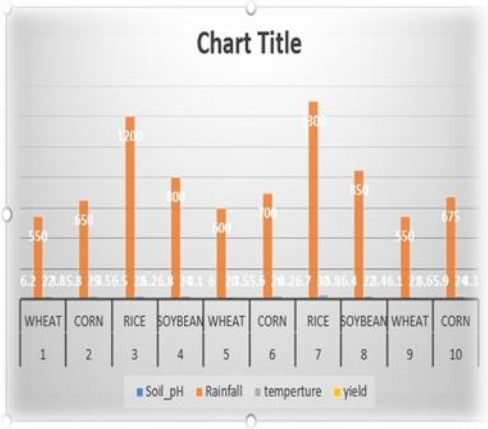

Fig: 5.2. The chart of the crops.

The figure 5.2 refers the chart of crops that contains wheat, corn, rice, etc…

VI. CONCLUSION

The study concludes that machine learning algorithms, including decision trees, random forests, and support vector machines, are effective tools for assessing plant suitability in diverse soil conditions. By incorporating key soil factors like pH, moisture content, and nutrient levels, these models can predict the optimal plant species for specific soil profiles with high accuracy. The research underscores the value of data-driven approaches in modern agriculture, highlighting how these techniques can support sustainable and efficient crop cultivation. The findings offer practical insights for farmers and agricultural experts, enabling more informed decision-making for improved productivity and sustainability.

VII. FUTURE ENHANCEMENT

Enhancing crop yield prediction through feature enrichment using machine learning techniques involves integrating diverse data sources and incorporating temporal, spatial, and domain-specific features. By leveraging multisource data such as remote sensing imagery, weather data, soil properties, and agronomic management practices, a more holistic understanding of the factors influencing crop yield can be attained. This approach allows for the extraction of spectral and vegetation indices from remote sensing data to characterize crop health and growth, while weather variables provide insights into climatic conditions impacting crop development. Incorporating temporal features like historical yield data and seasonal trends captures the dynamic nature of crop growth, while spatial features derived from geographic information system (GIS) data account for variations in soil properties and microclimate conditions. By enriching the feature space with relevant information, machine learning models can better capture the complex interactions between environmental factors and crop productivity, leading to improved accuracy in crop yield prediction.

REFERENCES

[1] P. Priya, U. MuthaiahM. Balamurugan. Predicting yield of the crop using machine learning algorithm. International Journal of Engineering Science Research Technology.

[2] J. Jeong, J. Resop, N. Mueller and team. Random forests for global and regional crop yield prediction. PLoS ONE Journal.

[3] Narayanan Balakrishnan and Dr.GovindarajanMuthukumarasamy.Crop production Ensemble Machine Learning model for prediction. International Journal of Computer Science and Software Engineering (IJCSSE).

[4] S. Veenadhari, Dr. Bharat Misra, Dr. CD Singh. Machine learning approach for forecasting crop yield based on climatic parameters. International Conference on Computer Communication and Informatics (ICCCI).

[5] S. ShaharBanu, "An Approach for identifying cancer using Support vector machine algorithm" in International Journal of Information Research and Review, April 2015, VOL2.ISS4, PP 660-663 ISSN 2349 91461.

[6] S. ShaharBanu,Dr.V.Saravanan,Dr.R.Sriram Published a paper "Extracting Peculiar data from Multidatabases Using Agent Mining" in International Journal of Recent Technology and Engineering (ISSN: 2277-3878, Volume-2, Issue-1,) (Impact Factor 1.0).

[7] 7.S.ShaharBanu,Dr.R.Srinivasan published a paper titled 'Peculiar data identification in multi databases using

RVER model' in International Journals of Technology and Engineering System,during 2010 (**ISSN :0976-269X)(Vol No1, page no 57-62)).**

[8] S. ShaharBanu,Dr,V. Saravanan published a paper titled "An Approach For Discovering Interesting Patterns In Multidatabases Using Peculiar Mining" to International Journal of Engineering Science Technology. (ISSN : 0975-5462 Vol. 3 No. 4)during April 2011 (**IC=3.14)**