Comparative Study of Parametric and Non-Parametric Models in Crop Recommendation

Krishna Ramakrishnan

Abstract

Agriculture is facing mounting challenges from population growth, food insecurity, soil degradation, and climate variability, making traditional practices insufficient to sustain production. AI and machine learning (ML) have recently been applied to agriculture, offering opportunities to improve crop selection and land use through data-driven recommendations. Previous studies have explored crop recommendation systems, but their findings have often been inconsistent due to differences in datasets and methods. This study addresses that gap by benchmarking parametric and non-parametric models on the same dataset to evaluate their performance in a multi-class crop recommendation task. Using an agricultural dataset of 8,000 entries across 11 crop types, data preprocessing included one-hot encoding, label encoding, and median imputation. A neural network implemented in TensorFlow was compared against K-Nearest Neighbors (KNN) and Random Forest models. Results showed that the neural network achieved 80–83% accuracy, while the non-parametric models remained near 9–10%, close to random guessing. These findings suggest that parametric models are better suited for capturing the complex, non-linear patterns in agricultural data. Despite limitations such as overfitting and reliance on a single dataset, the study highlights the potential of AI to provide more accurate and sustainable crop recommendations, ultimately supporting global food security.

Introduction

Agriculture has remained one of the most essential sectors for sustaining human life, yet it has faced increasing challenges in recent decades. With the global population continuing to rise and food insecurity becoming a growing issue, farmers must meet higher production quotas while dealing with environmental pressures, soil degradation, and unpredictable climate conditions. Traditional farming practices have been proven insufficient to address these complex issues, even if they are the foundation of agriculture. In recent years, technological innovations have emerged to support agriculture, including advanced sensors for monitoring soil and crop health, innovations in crop engineering, and software systems for optimizing production. These advancements have started a shift toward smart agriculture, where data-driven decisions have improved efficiency, reduced waste, and optimized yields. Such methods have been crucial in supporting sustainable food production and meeting the demands of a rapidly changing world.

In the current state of our world, populations are constantly growing, food insecurity is rising and the agricultural sector is struggling. Several fields, including finance, healthcare, e-commerce, and many more have been revolutionized by AI. Farmers throughout the world are struggling due to problems such as selecting the best crops for their soil and climate, optimizing land use, and adapting to changing local conditions. As a result, the implementation of AI and data science has become crucial to increase efficiency and meet global food demand. AI and ML-based crop recommendation systems have the potential to address these issues by analyzing large datasets to generate accurate and even location-specific recommendations. While prior studies have used machine learning to create crop recommendation systems, their results have been scattered and inconsistent. This study aimed to address this gap by

comparing the performance of parametric models and non-parametric models on a uniform dataset. The specific objective was to observe differences in performance between a neural network and two non-parametric models (K-Nearest Neighbors and Random Forest) to evaluate their suitability for a multi-class crop recommendation task.

Materials and methods

1. Data Preparation

The dataset used for this project was publicly available on Kaggle and is licensed under MIT (https://www.kaggle.com/datasets/shankarpriya2913/crop-and-soil-dataset). It contains distinct soil and environmental input features including temperature, humidity, moisture, soil type, nitrogen, phosphorus, potassium, and fertilizer used. There are 8000 data points in this dataset with 11 different crop classes.

2. Data Preprocessing

Data preprocessing followed conventional standards in the field. The dataset was loaded using the Pandas library, and the input features were separated from the target variable (crop type). All categorical features were converted into numerical format using one-hot encoding, while the target variable was converted into integer labels using scikit-learn's LabelEncoder. Any missing values were filled with the median of their respective columns. The dataset was split into training and testing sets in a 70:30 ratio.

3. Parametric Models

A neural network was implemented using Keras, the API for TensorFlow. The model consisted of three hidden layers, each using the ReLU activation function. Batch normalization was applied between layers to improve training speed. The output layer used a softmax activation function to classify the data into 11 crop types. The model was compiled with the Adam optimizer and a Sparse Categorical Cross-Entropy loss function. Training was done over 2,500 epochs with a batch size of 64 and a validation split of 20%. Model performance was measured using accuracy, calculated by comparing predicted class labels with the true labels in the testing dataset.

4. Non-Parametric Models

The K-Nearest Neighbors model was implemented using scikit-learn's KNeighborsClassifier, with an initial hyperparameter of 8 neighbors. A hyperparameter optimization loop was created to test values of k from 1 to 100, recording accuracy for each. The Random Forest model was implemented using scikit-learn's RandomForestClassifier with 91 trees as the main hyperparameter. Similar to KNN, an optimization loop was used to find the number of trees with the best accuracy. Both models were trained on the training dataset and tested on the testing

dataset. Accuracy was calculated by comparing predicted labels to actual labels. To demonstrate model predictions, a single sample from the test set was classified and the result was converted back to its corresponding crop name.

***Results*:**

EDA

The dataset used this study consists of 8000 records with a mix of numerical and categorical features important to agricultural conditions, including Temperature, Humidity, Moisture, Nitrogen, Phosphorous, Potassium, Soil Type, and Fertilizer Name, with Crop Type as the target variable. To keep the data clean and usable, all numerical columns were checked and converted to the correct format, and any missing values were filled in using the median value of each column. Categorical columns were encoded using one-hot encoding for model training and label encoding for visualizations.

A pair plot of all features (Figure 5) revealed heavy overlap among crop classes, especially in features related to soil nutrients and environmental conditions. There were no clear linear separations between classes, showing the need for complex, non-linear models to effectively capture hidden patterns. The correlation matrix (Figure 6) confirmed that most features are weakly correlated with each other, with the exceptions of a moderate positive correlation between Temperature and Humidity ($r = 0.53$), and a strong negative correlation between Nitrogen and Phosphorus ($r = -0.64$). Because every feature adds something unique, all 8 can be kept without worrying about redundant information.

Figure 5: This pair plot shows the relationships between different factors like temperature, humidity, and nutrients in the soil. The plots along the diagonal show how each individual factor is spread out. The other plots show how two different factors relate to each other. The different colors represent various crop types, which helps to see how these relationships change depending on the crop.

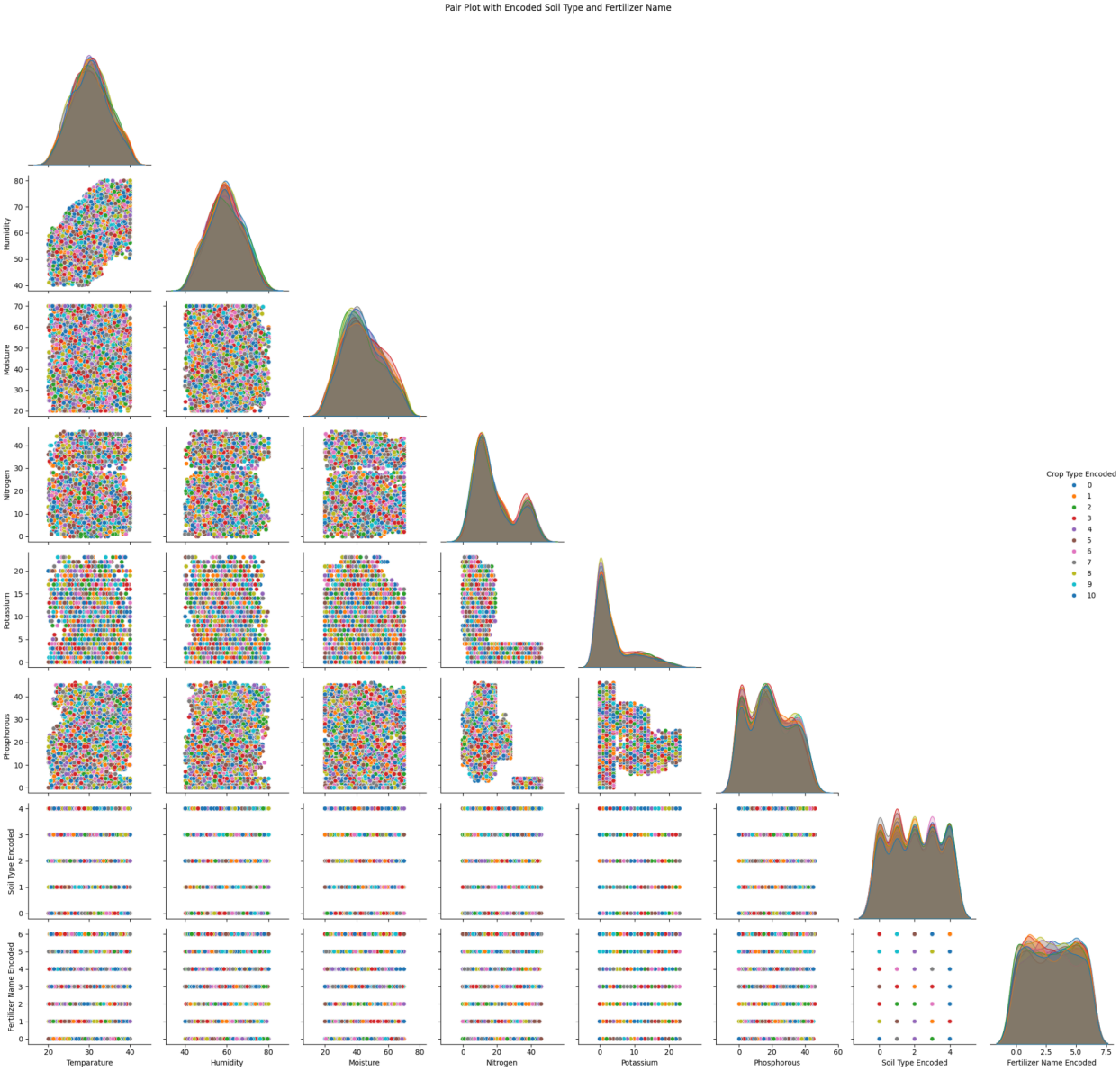
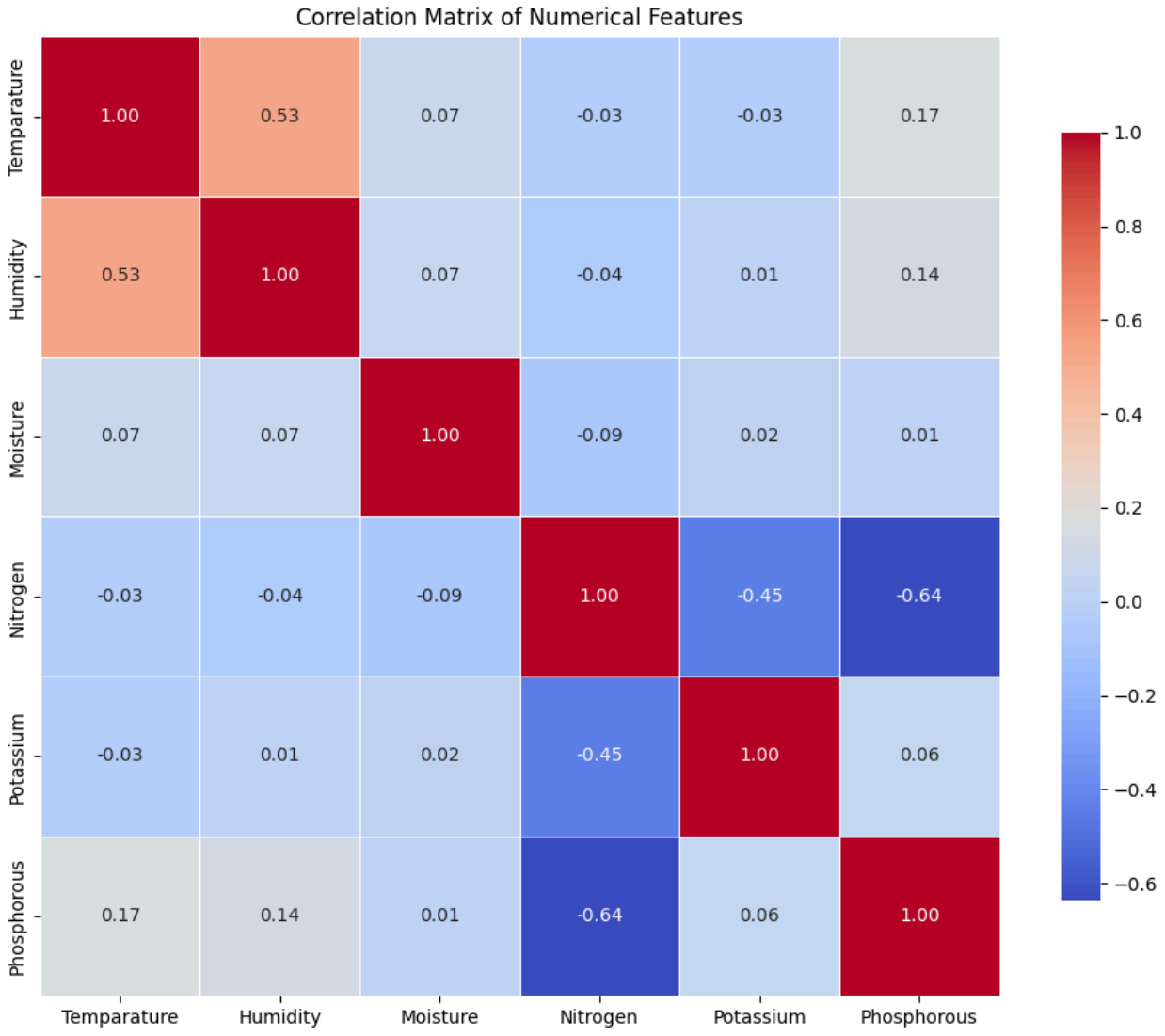
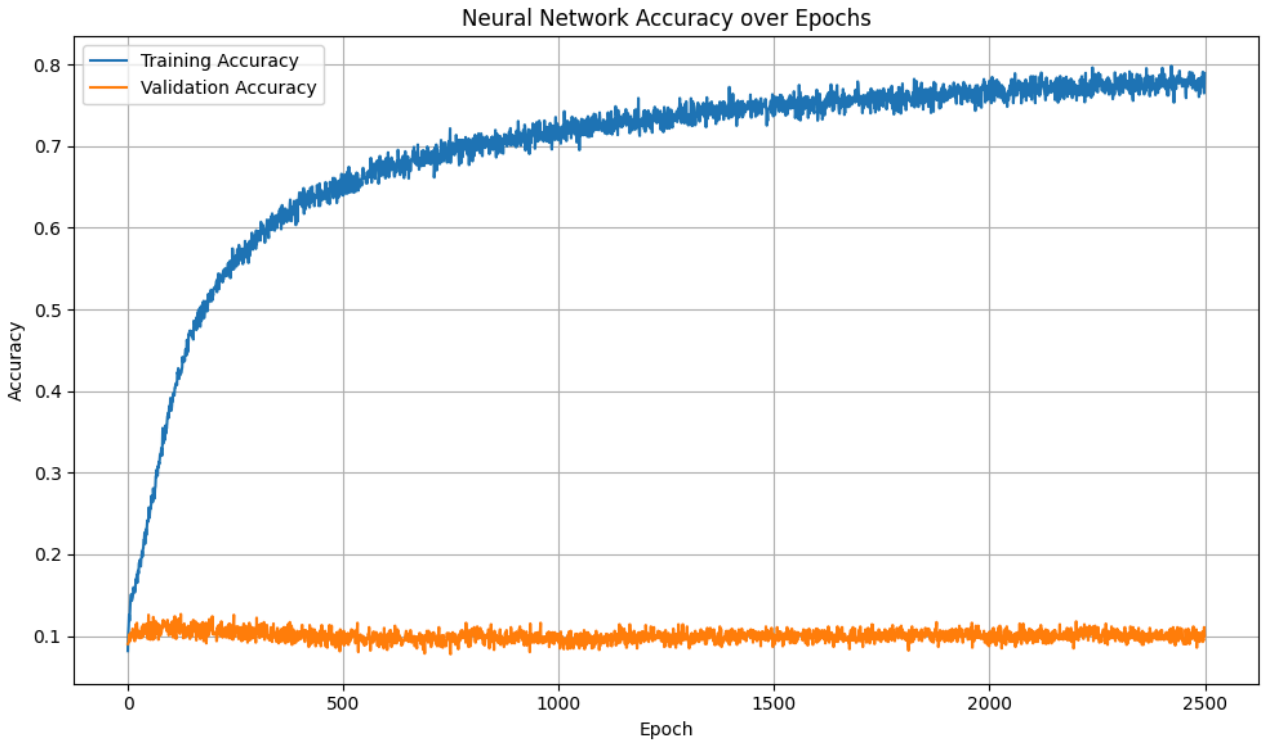


Figure 6: This matrix shows how the different features relate to each other. Temperature and Humidity have a strong positive relationship, meaning they tend to increase or decrease together. Nitrogen has a negative relationship with Phosphorus and Potassium, suggesting that when nitrogen levels are high, the others tend to be low.



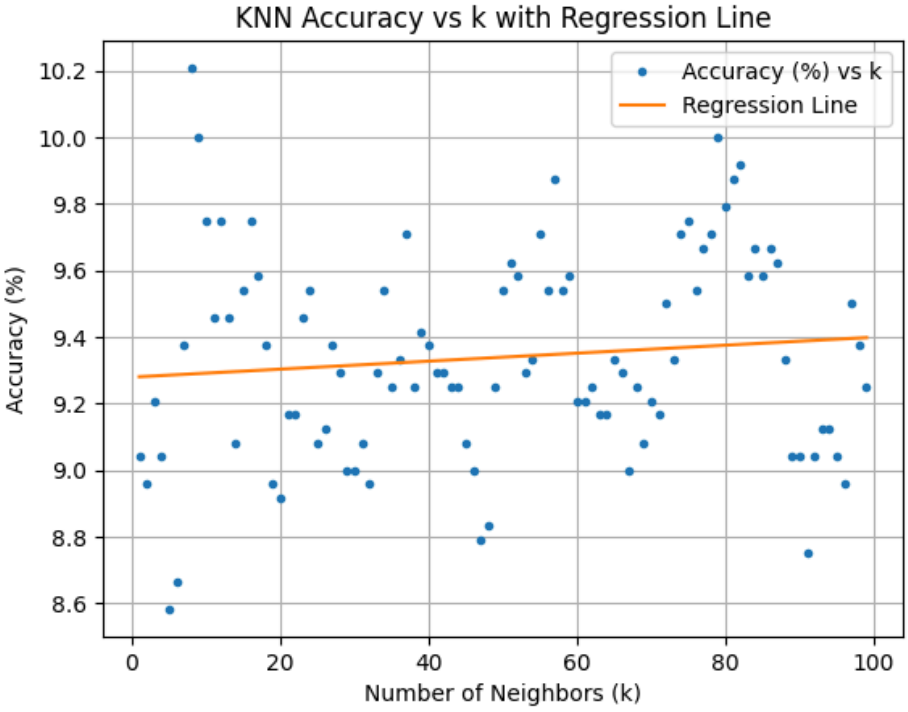
The neural network achieved an accuracy score of 80% around the 2000th epoch and fluctuated between 80 to 83% for the remaining epochs (Figure 1). However, the validation accuracy consistently stayed at 9 to 11%, indicating that there was significant overfitting of the data (about the same accuracy as random guessing). Loss continued to decrease until the model converged. It was able to classify the majority of the crop types but failed in cases where 2 crop types had very similar ideal conditions.

Figure 1: Neural network training and validation accuracy over 2500 epochs. The neural network with 3 hidden layers was trained on soil and environmental data (8000 data points, 8 features, 11 crop types in the target variable). Training accuracy (blue) increased steadily over time until it plateaued at 80-83% accuracy after the 2000th epoch. Validation accuracy (orange) remained near 10%, indicating overfitting.



The KNN model achieved a max accuracy score of 10.2% (at k = 8) across all tested values of k from 1-100. The lowest accuracy score of 8.58% occurs at k = 5. The regression line in the graph between k values and accuracy (Figure 4) shows that there is almost no correlation between the 2 (correlation coefficient of 0.11).

Figure 4: This graph shows the model's accuracy as we increased the number of trees it used. The accuracy went up and down, peaking around 9.4%. After about 100 trees, the accuracy became more stable, but didn't consistently improve.



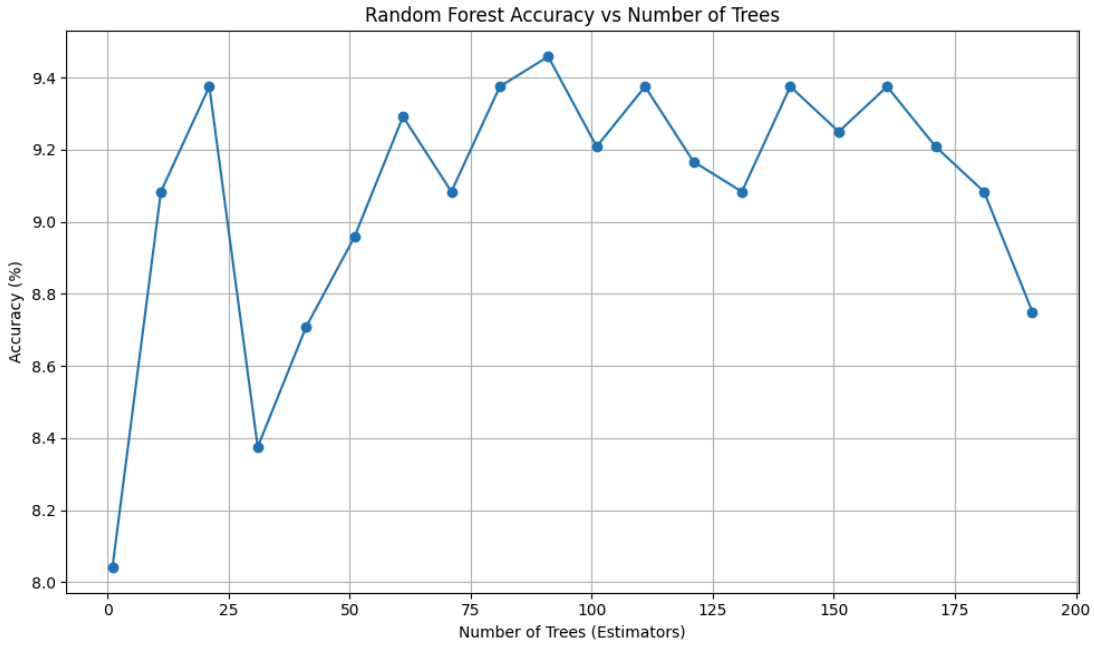
Discussion

This study explored the role of artificial intelligence and machine learning in agriculture, focusing on the development of a crop recommendation system. Agriculture today faces immense challenges such as soil degradation, unpredictable climate, and the need for higher yields to meet food demands. AI-driven approaches enable the processing of large datasets and provide precise crop recommendations that traditional methods cannot achieve. To evaluate this potential, three models were selected for comparison: a neural network representing parametric methods, and K-Nearest Neighbors (KNN) and Random Forest representing non-parametric methods. Using a standardized dataset of 8,000 entries containing soil and environmental features, each model was trained and tested on the same split of data.

By comparing the parametric approach (neural network) with the non-parametric models (KNN and Random Forest), we directly compared different machine learning models on a multi-class crop classification task in agriculture. The neural network achieved 80% accuracy around the 2000th epoch and maintained a range of 80% to 83% for the remainder of the training (Figure 1). This indicated good generalization and model stability. On the other hand, the non-parametric models performed poorly. The KNN model, with a hyperparameter of 70 neighbors, achieved an accuracy of only 9.21% (Figure 4), and testing various values of k from 1 to 100 did not improve the results. Similarly, the Random Forest classifier, using 100 trees, also achieved an accuracy of just 9%, showing limited promise (Figure 3). These results suggest that the parametric neural network may be a better option for capturing the hidden

patterns and complexities in the dataset, while the non-parametric models struggled due to high dimensionality in the data set, or due to difficulties in the multiclass classification task. Overall, the findings support that parametric models such as neural networks perform better in a multiclass application such as a crop recommendation system.

Figure 3: This graph shows the model's accuracy as we increased the number of trees it used. The accuracy went up and down, peaking around 9.4%. After about 100 trees, the accuracy became more stable, but didn't consistently improve.



These findings indicate that parametric models such as neural networks are better suited for multi-class agricultural recommendation systems, as they can capture non-linear patterns in high-dimensional data. However, the neural network's overfitting suggests that more thorough preprocessing, extra techniques to reduce overfitting and improve generalization, or access to larger and more diverse datasets would be necessary to improve generalization. If more time and resources were available, the study could be extended by testing other advanced architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), or transformers, and by including real-world field data beyond the Kaggle dataset. Limitations of this study include reliance on a single dataset, limited hyperparameter exploration, and a lack of evaluation using domain-specific metrics such as crop yield efficiency or economic viability. Despite these constraints, the work highlights the potential of AI in agriculture. If fully fleshed out, an accurate crop recommendation system could significantly reduce resource waste, improve food security, and provide small-scale farmers with personalized insights that directly impact their way of life and contribute to sustainable agricultural practices worldwide.

References

Shankar. (2025, January 28). *Crop and soil dataset*. Kaggle. https://www.kaggle.com/datasets/shankarpriya2913/crop-and-soil-dataset

Waikar, V., Thorat, S., Ghute, A., Rajput, P., & Shinde, M. (n.d.). Crop Prediction based on Soil Classification using Machine Learning with Classifier Ensembling. In *International Research Journal of Engineering and Technology*. Retrieved August 24, 2025, from https://www.irjet.net/archives/V7/i5/IRJET-V7I5931.pdf

scikit-learn. (2019). *sklearn.neighbors.KNeighborsClassifier — scikit-learn 0.22.1 documentation*. Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

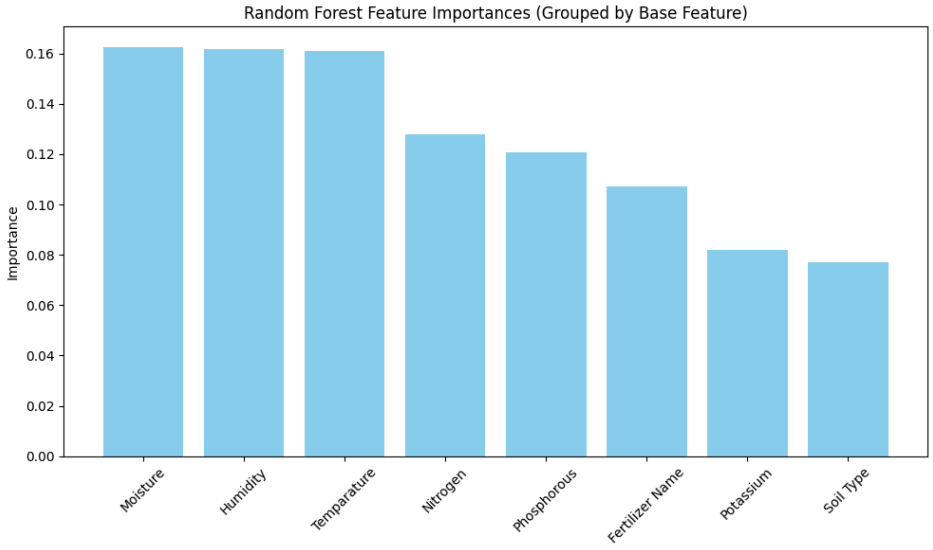
Hardesty, L. (2017, April 14). *Explained: Neural networks*. MIT News; Massachusetts Institute of Technology. https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

Kumar, R. (2024, May 20). *Batch Normalization: Theory and TensorFlow Implementation*. Datacamp.com; DataCamp. https://www.datacamp.com/tutorial/batch-normalization-tensorflow

Shields, R. (2024). *The Use of Ai (Artificial Intelligence) in Agriculture & Farming*. Www.agrirs.co.uk. https://www.agrirs.co.uk/blog/2024/02/the-use-of-ai-artificial-intelligence-in-agriculture-and-farming?source=google.com

Other Tables & Figures

Figure 2: This chart shows which features were most important for the Random Forest model's predictions. The model relied most heavily on Moisture, Humidity, and Temperature, while Potassium and Soil Type were the least important.


Random Forest Feature Importances (Grouped by Base Feature)
Importance
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
Moisture
Humidity
Temparature
Nitrogen
Phosphorous
Fertilizer Name
Potassium
Soil Type