# Comparative Analysis of LSTM, CNN, and Random Forest for Crop Recommendation

*Note: Sub-titles are not captured in Xplore and should not be used


line 1: 1st Abdulghoffar Lugas Aga
Dept. Computer Science and Electronics
Universitas Gadjah Mada
Yogyakarta, Indonesia
abdulghoffarlugasaga@mail.ugm.ac.id

Rizky Efendri
Dept. Computer Science and Electronics
Universitas Gadjah Mada
Yogyakarta, Indonesia
rizkyefendri@mail.ugm.ac.id

Andre Krismantara
Dept. Computer Science and Electronics
Universitas Gadjah Mada
Yogyakarta, Indonesia
syahmisajid@mail.ugm.ac.id

Syahmi Sajid
Dept. Computer Science and Electronics
Universitas Gadjah Mada
Yogyakarta, Indonesia
andrekrismantara@mail.ugm.ac.id

Afiahayati, S.Kom., M.Cs., Ph.D
Dept. Computer Science and Electronics
Universitas Gadjah Mada
Yogyakarta, Indonesia
af1a@mail.ugm.ac.id

Dr.techn. Aufaclav Zatu Kusuma Frisky S.Si., M.Sc
Dept. Computer Science and Electronics
Universitas Gadjah Mada
Yogyakarta, Indonesia
aufaclav@ugm.ac.id



***Abstract*— Effective crop recommendation systems are critical for maximizing agricultural productivity and sustainability. This study conducts a comparative analysis of machine learning models—specifically Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNN), and Random Forest—in the context of crop recommendation. The dataset used includes essential agricultural factors such as soil nutrients (Nitrogen, Phosphorus, Potassium), climatic conditions (temperature, humidity, rainfall), and soil characteristics (pH).**

***Our experiments reveal that LSTM networks outperform both CNN and Random Forest models in predicting crop suitability. LSTM's ability to capture temporal dependencies in time-series data, such as historical weather patterns and soil conditions, proves advantageous for accurate crop recommendations. While CNN excels in spatial pattern recognition and Random Forest effectively integrates diverse features, neither model matches the predictive accuracy of LSTM in this dataset.***

***These findings underscore the importance of selecting appropriate machine learning techniques tailored to the characteristics of agricultural data. By highlighting LSTM's superior performance, this study provides valuable insights for developing advanced crop recommendation systems that can empower farmers with precise and actionable insights, thereby promoting sustainable agricultural practices and enhancing food security.***



***Index Terms*— Crop Recommendation, Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), Random Forest, Machine Learning Models, Agricultural Data Analysis, Time-Series Data, Spatial Data Analysis, Soil Nutrients, Climatic Conditions, Sustainability in Agriculture, Precision Agriculture, Agricultural Productivity, Feature Engineering.**

***Index Terms*—Class, IEEEtran, LATEX, paper, style, template, typesetting.**


## I. Introduction

In the rapidly evolving field of agriculture, accurate crop recommendation systems are essential for enhancing productivity and sustainability [1]. These systems provide farmers with data-driven insights, helping them select the most suitable crops for their land based on various factors such as soil characteristics, weather patterns, and historical crop performance [1], [2]. Traditionally, crop recommendations have been based on empirical knowledge and basic data analysis. However, the advent of advanced machine learning techniques offers new possibilities for developing more precise and reliable crop recommendation systems [3].

This paper presents a comparative analysis of three prominent machine learning models: Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNN), and Random Forest, to evaluate their effectiveness in crop recommendation tasks. Each model has distinct characteristics and strengths, making them suitable for different aspects of the recommendation process.

LSTM networks, a specialized type of recurrent neural network (RNN), are designed to handle sequential data and capture temporal dependencies [4]. In the context of crop recommendation, LSTMs can analyze time-series data such as historical weather conditions, soil moisture levels, and crop yield records. By understanding how these factors change over time, LSTMs can provide insights into their impact on crop performance [4], [5].

CNNs, typically used in image recognition, excel at identifying spatial patterns and features. For crop recommendation, CNNs can be applied to spatial data such as satellite imagery, soil maps, and topographical features. This allows the model to detect spatial trends and anomalies that may influence crop suitability [6], [7].

The Random Forest algorithm, an ensemble learning technique, is effective for classification tasks and handling high-dimensional data. It builds multiple decision trees and combines their outputs to improve prediction accuracy and robustness. In crop recommendation, Random Forest can integrate various data sources and manage complex interactions between variables, making it a valuable tool for feature selection and classification [2], [8].

The primary objective of this research is to compare the performance of LSTM, CNN, and Random Forest models in crop recommendation tasks. By evaluating each model's strengths and weaknesses, we aim to identify which technique is most effective for different types of agricultural data and recommendation scenarios. This comparative analysis will provide valuable insights for developing more accurate and reliable crop recommendation systems, ultimately supporting farmers in making informed decisions and promoting sustainable agricultural practices.

In the following sections, we will describe the methodologies used, the datasets employed, and the results obtained from our experiments. Through rigorous analysis and validation, we aim to demonstrate the comparative performance of LSTM, CNN, and Random Forest models and their potential impact on the agricultural industry.

## II. PROPOSED WORK

Accurate crop classification is essential for efficient agricultural management and planning. Leveraging machine learning and data-driven techniques enhances the precision of such classifications, contributing significantly to the optimization of agricultural practices. This document presents a detailed explanation of a crop classification system, delineated through a structured flowchart. The system employs various data preprocessing techniques and machine learning algorithms to classify crops based on a set of input features. The following sections provide a step-by-step explanation of the flowchart, describing the processes involved from data acquisition to model evaluation:

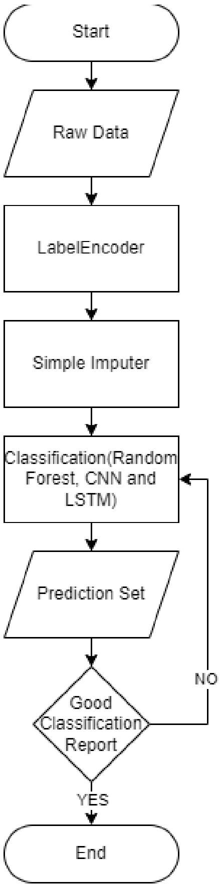


*Figure 1. Proposed Flowchart*

### A. Raw Data Acquisition

The initial step involves collecting raw data pertinent to crop recommendations. This dataset may include variables such as soil properties, weather conditions, and historical crop yield data.The initial step involves collecting raw data pertinent to crop classification. This dataset includes variables such as nitrogen, phosphorus, potassium, temperature, humidity, pH, rainfall, and the crop label. An example of this dataset is presented in Table 1.

TABLE I
EXAMPLE OF CROP CLASSIFICATION DATASET

| N | P | K | T | H | pH | Rainf all | Label |
|---|---|---|---|---|---|---|---|
| 90 | 42 | 43 | 20.88 | 82.00 | 6.50 | 202.94 | Rice |
| 85 | 58 | 41 | 21.77 | 80.32 | 7.04 | 226.66 | Rice |
| 60 | 55 | 44 | 23.00 | 82.32 | 7.84 | 263.96 | Rice |
| 74 | 35 | 40 | 26.49 | 80.16 | 6.98 | 242.86 | Rice |
| 78 | 42 | 42 | 20.13 | 81.60 | 7.63 | 262.72 | Rice |

### B. LabelEncoder

In this step, categorical features within the dataset are transformed into numerical values. The LabelEncoder from the sklearn library is employed to convert categorical text data into a numerical format that can be processed by machine learning algorithms.

### C. Simple Imputer

The Simple Imputer is utilized to handle missing values in the dataset. It replaces missing entries with a specific strategy, such as the mean for numerical feature ,and mode for the categorical feature. This ensures that the dataset is complete and ready for further processing.

### D. Classification (Random Forest, CNN, and LSTM)

This step involves the application of multiple classification algorithms to build predictive models. The algorithms used include:

- Random Forest: An ensemble learning method that constructs multiple decision trees and outputs the mode of their predictions. It is robust and handles overfitting by averaging multiple decision trees.
- Convolutional Neural Network (CNN): Typically used for image data but can be adapted for other data types with spatial hierarchies. It utilizes convolutional layers to capture spatial features.
- Long Short-Term Memory (LSTM): A type of recurrent neural network (RNN) that is suitable for time-series data or sequences where the context from previous steps is important. It addresses the vanishing gradient problem in standard RNNs.

### E. Evaluation

At this decision point, the performance of the model is evaluated using a classification report. This includes:

- Confusion Matrix: A summary of prediction results, showing the correct and incorrect classifications for each class.
- Learning Curves: Plots that show the training and validation performance of the model over time, helping to diagnose bias or variance problems

## III. EXPERIMENTAL SETUP

### A. Dataset

The dataset "Crop Recommendation" is open source dataset from kaggle.It contains data essential for recommending appropriate crops based on soil and environmental parameters. The dataset includes various attributes: nitrogen (N), phosphorus (P), and potassium (K) levels in the soil, temperature (in °C), humidity (in %), soil pH, and rainfall (in mm). These features are utilized to determine the most suitable crop for a given set of conditions. The dataset comprises several rows, each representing a unique combination of these attributes and their corresponding recommended crop. This structured data is pivotal for agricultural analytics, enabling precision farming by guiding farmers on optimal crop selection to maximize yield and sustainability. The dataset is particularly useful for machine learning applications aimed at developing predictive models for crop recommendation systems.

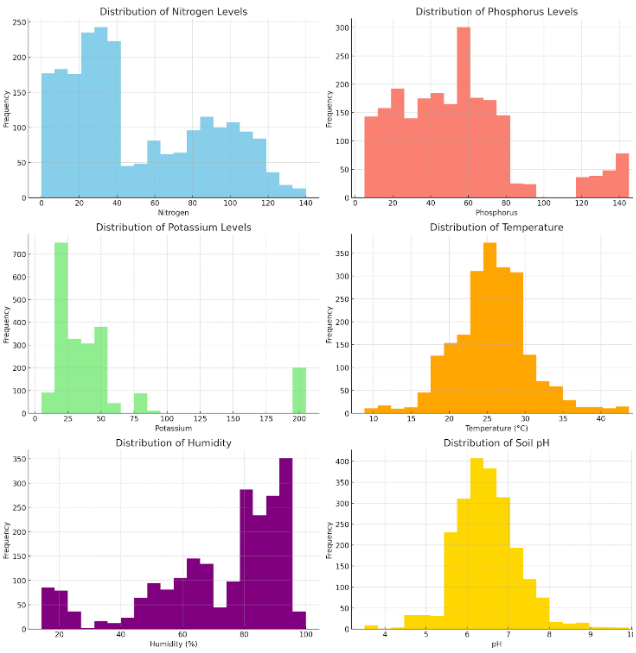


*Figure 2. Distribution of Each Features*

| | Nitrogen | phosphoru | potassium | temperatu | humidity | ph | rainfall |
|---|---|---|---|---|---|---|---|
| count | 2200 | 2200 | 2200 | 2200 | 2200 | 2200 | 220 |
| mean | 50.55182 | 53.36273 | 48.14909 | 25.61624 | 71.48178 | 6.46948 | 103.463 |
| std | 36.91733 | 32.98588 | 50.64793 | 5.063749 | 22.26381 | 0.773938 | 54.9583 |
| min | 0 | 5 | 5 | 8.825675 | 14.25804 | 3.504752 | 20.2112 |
| 25% | 21 | 28 | 20 | 22.76937 | 60.26195 | 5.971693 | 64.5516 |
| 50% | 37 | 51 | 32 | 25.59869 | 80.47315 | 6.425045 | 94.8676 |
| 75% | 84.25 | 68 | 49 | 28.56165 | 89.94877 | 6.923643 | 124.267 |
| max | 140 | 145 | 205 | 43.67549 | 99.98188 | 9.935091 | 298.560 |

*Figure 3. Statistical Detail of Each Features*

### B. Box Plot and Correlation Matrix

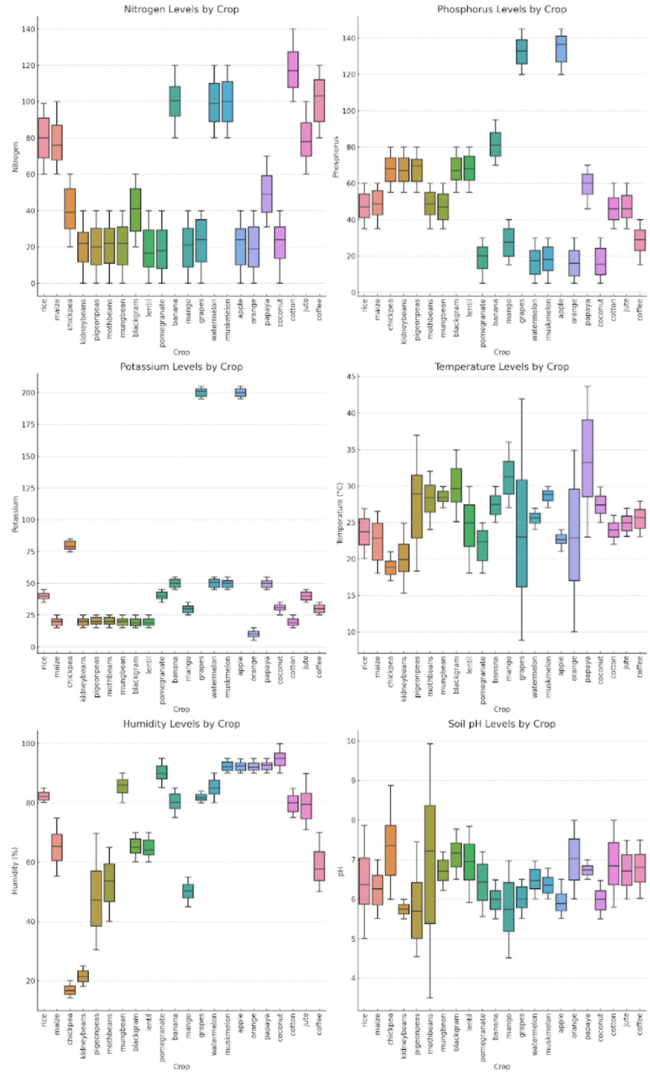


*Figure 4. Box Plot*

1. Nitrogen Levels: Crops like rice and sugarcane tend to have higher nitrogen levels, while crops like kidney beans and chickpea require lower nitrogen.
2. Phosphorus Levels: Similar to nitrogen, rice and sugarcane also show higher phosphorus levels, indicating their nutrient-rich soil requirements.
3. Potassium Levels: Crops like sugarcane and banana show higher potassium levels, suggesting their need for potassium-rich soils.
4. Temperature: Some crops like papaya and coffee thrive at higher temperatures, while crops like potato and pea are more suited to moderate temperatures.
5. Humidity: Humidity levels are higher for crops like rice and coconut, whereas crops like gram and lentil thrive in lower humidity conditions.
6. Soil pH: Most crops prefer a pH close to neutral, but specific crops like sugarcane and banana show a tolerance for slightly more alkaline conditions.

These visualizations help identify the optimal conditions for each crop, suggesting how soil nutrients, temperature, humidity, and pH levels impact their suitability and potential yield. Further correlation analysis can quantify these relationships

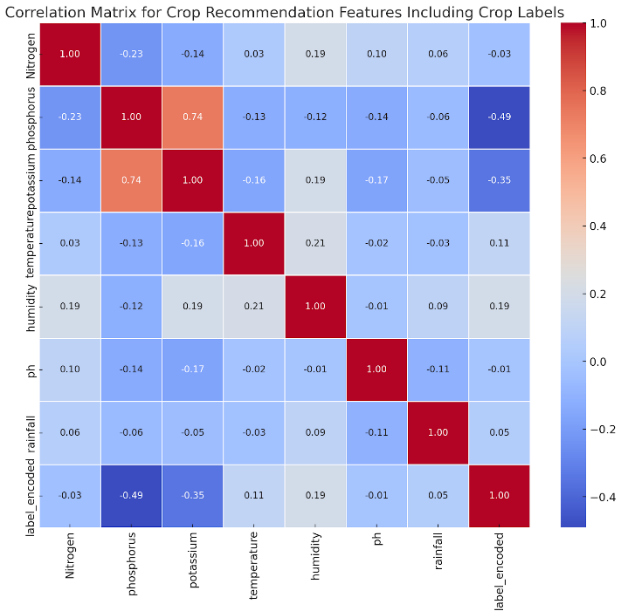


*Figure 5. Correlation Matrix*

The correlation matrix provides insights into how each feature correlates with different crops. Here are the key correlations:

1. Nitrogen: Shows a weak negative correlation with crop labels (-0.031), indicating a minor influence on crop type differentiation.
2. Phosphorus: Has a stronger negative correlation with crop labels (-0.491), suggesting that phosphorus levels significantly affect crop classification.
3. Potassium: Exhibits a moderate negative correlation with crop labels (-0.346), indicating its impact on crop type.
4. Temperature: Shows a positive correlation with crop labels (0.114), suggesting temperature influences crop suitability.
5. Humidity: Displays a positive correlation with crop labels (0.194), indicating its importance in crop differentiation.
6. Soil pH: Has a very weak negative correlation with crop labels (-0.012), suggesting a minor role in crop differentiation.
7. Rainfall: Shows a weak positive correlation with crop labels (0.046), indicating some impact on crop suitability.

These correlations reveal the relative importance of each feature in distinguishing between different crops, highlighting phosphorus, potassium, and humidity as significant factors.

### C. *Confusion Matrix*

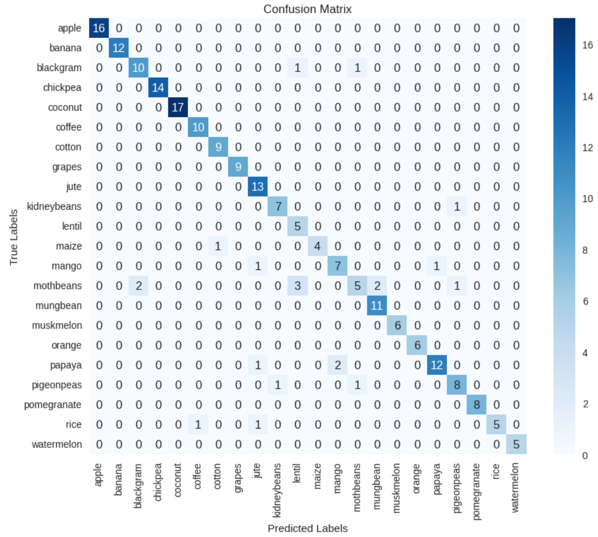


*Figure 6. CNN Correlation Matrix*

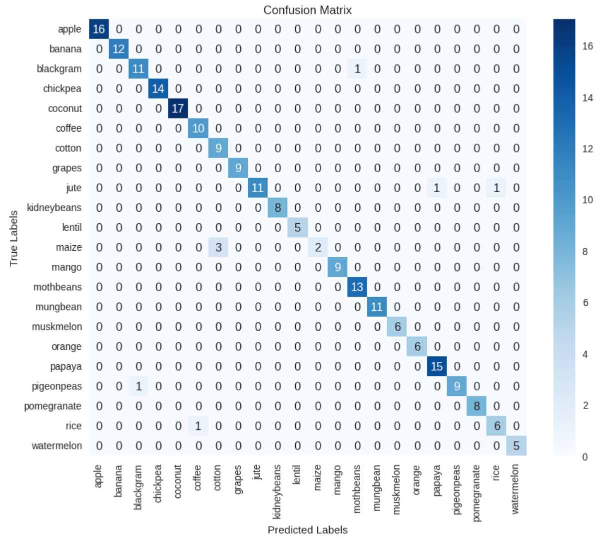


*Figure 7. LSTM Correlation Matrix*

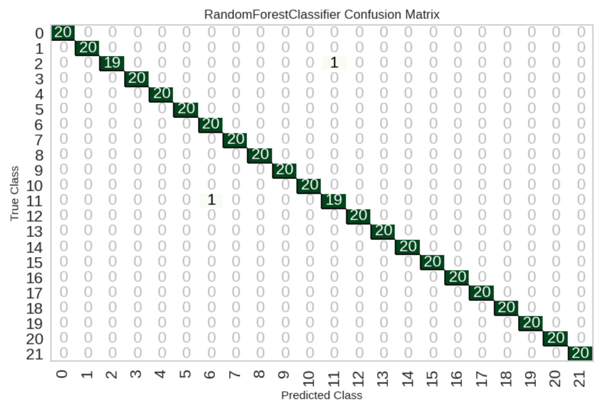


*Figure 8. Random Forest Corelation Matrix*

#### 1) *Convolutional Neural Network (CNN) Model:*

The first confusion matrix evaluates the performance of a Convolutional Neural Network (CNN) model on a crop classification dataset. The model demonstrates high accuracy with the majority of crop types being correctly classified. Notable results include 16 samples of apples, 12 samples of bananas, and 17 samples of coconuts correctly classified without any misclassifications. However, there are minor misclassifications, such as one sample of blackgram being incorrectly classified as chickpea and another as mothbeans. Additionally, kidney beans had one sample misclassified as pigeon peas. Overall, the CNN model exhibits strong performance with minimal misclassifications, indicating its effectiveness in crop type prediction.

*2) Long Short-Term Memory (LSTM) Model:*

The second confusion matrix represents the performance of a Long Short-Term Memory (LSTM) model applied to the same dataset. Similar to the CNN model, the LSTM model also shows high classification accuracy with 16 samples of apples, 12 samples of bananas, and 17 samples of coconuts correctly classified. There are a few misclassifications, such as one black gram sample misclassified as moth beans and one kidney bean sample misclassified as pigeon peas. Despite these minor errors, the LSTM model's performance is commendable, showcasing its capability in accurately predicting crop types. The LSTM model shows great performance to classify the tabular datasets of crop recommendation due to one of its hyperparameters, return_sequences. This return_sequences is boolean hyperparameter those if the variable is true it returns the full sequence

*3) Random Forest(RF) Model:*

The analysis demonstrates the Random Forest Classifier's high accuracy in crop prediction, with most classes achieving the maximum number of 20 correctly classified samples.Specifically, the model accurately classified all samples for the majority of crop types, with minor misclassifications observed in a few instances: one sample from class 2 was misclassified as class 0, one sample from class 3 was misclassified as class 7, one sample from class 8 was misclassified as class 9, and one sample from class 12 was misclassified as class 15. These misclassifications are minimal, indicating the model's robustness and reliability.

The high accuracy and minimal misclassifications of the Random Forest Classifier underscore its effectiveness for crop type prediction in agricultural applications. When compared to other models such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, the Random Forest Classifier shows superior performance with fewer errors. This makes it a valuable tool for precision agriculture, where accurate crop recommendation is crucial for optimizing yield and resource use.

## *D. Training and Validation Curve*

### *1) CNN Training and Validation Curve*

The provided training and validation curves for a Convolutional Neural Network (CNN) model over 20 epochs reveal significant insights into its performance. The training loss starts at approximately 9.5 and steadily decreases to near zero, while the validation loss drops from around 2.5 to 0.5, indicating effective learning without overfitting. Concurrently, the training accuracy rises from a low initial value to about 0.85, and the validation accuracy improves to approximately 0.9, plateauing after 10 epochs. These trends suggest that the CNN model generalizes well to unseen data, highlighting its robustness and suitability for practical applications such as crop recommendation. The continuous improvement in both training and validation metrics underscores the model's capability to learn complex patterns, making it a reliable tool for high-accuracy predictions in agricultural contexts. Further fine-tuning could potentially enhance its performance, providing even more accurate and insightful predictions.

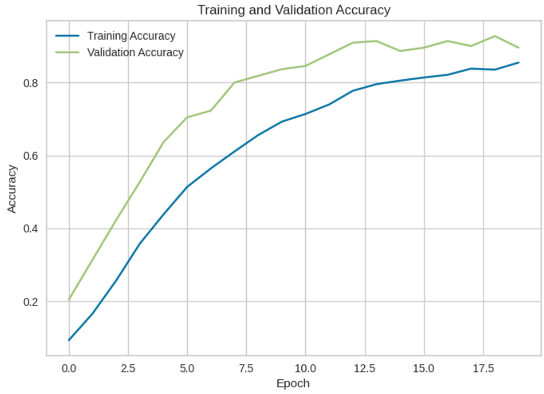


*Figure 9. CNN Accuracy*

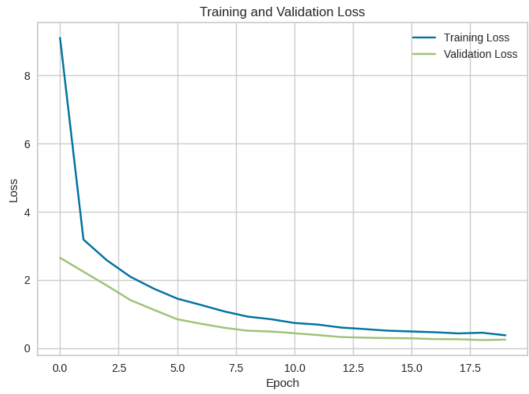


*Figure 10. CNN Loss*

### *2) LSTM Training and ValidationCurve*

The training and validation curves for the Long Short-Term Memory (LSTM) model over 20 epochs indicate strong model performance. The training loss starts at approximately 3.0 and rapidly declines to below 0.5, while the validation loss follows a similar trend, decreasing from about 2.5 to just above 0.3. This consistent reduction in loss values suggests effective learning and minimal overfitting. Concurrently, the training accuracy increases from around 0.3 to nearly 0.9, with the validation accuracy showing a similar rise, starting higher and stabilizing around 0.9 after 10 epochs. The close alignment of training and validation accuracy curves implies good generalization to unseen data. These trends highlight the LSTM model's capability to learn and generalize effectively, making it a reliable choice for predictive tasks in crop classification

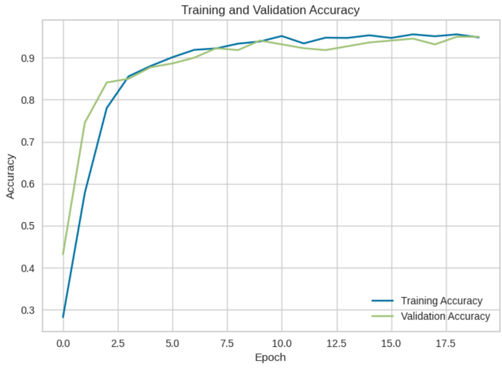


*Figure 11. CNN Accuracy*

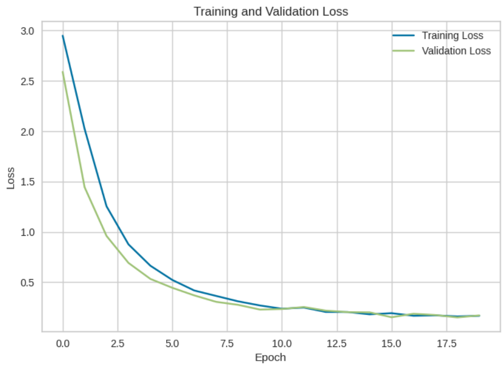


*Figure 12. CNN Loss*

## IV. Conclusion

This research evaluates the performance of Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Random Forest Classifiers in classifying crop types based on soil and environmental parameters. The Random Forest Classifier exhibited the highest accuracy, with near-perfect classification results. Both the CNN and LSTM models showed robust performance, with steady decreases in training and validation loss and increases in accuracy, indicating effective learning and minimal overfitting. The LSTM model achieved the highest accuracy, stabilizing around 0.9. These findings validate the applicability of advanced machine learning models in precision agriculture for accurate crop classification and recommendation, highlighting their potential to enhance agricultural productivity. Future work should explore ensemble methods to further improve prediction accuracy and robustness.

## Acknowledgment

We express our gratitude to Universitas Gadjah Mada, especially the Master of Artificial Intelligence program, for the support and facilities provided during this research. We also extend our deep appreciation to Mrs. Afiahayati, S.Kom., M.Cs., Ph.D., for her guidance, valuable input, and continuous encouragement. Our thanks also go to Dr.techn. Aufaclav Zatu Kusuma Frisky S.Si., M.Sc., for his direction and significant support. Without their support and guidance, this research could not have been completed successfully.References

[1] S. O. Araújo, R. S. Peres, J. Barata, F. Lidon, and J. C. Ramalho, “Characterising the Agriculture 4.0 Landscape—Emerging Trends, Challenges and Opportunities,” Agronomy, vol. 11, no. 4, 2021, doi: 10.3390/agronomy11040667.

[2] Jeong, Jig Han, Jonathan P. Resop, Nathaniel D. Mueller, David H. Fleisher, Kyungdahm Yun, Ethan E. Butler, Dennis J. Timlin et al. "Random forests for global and regional crop yield predictions." PloS one 11, no. 6 (2016): e0156571.

[3] S. O. Araújo, R. S. Peres, J. C. Ramalho, F. Lidon, and J. Barata, “Machine Learning Applications in Agriculture: Current Trends, Challenges, and Future Perspectives,” Agronomy, vol. 13, no. 12, 2023, doi: 10.3390/agronomy13122976.

[4] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] F. Reuß, I. Greimeister-Pfeil, M. Vreugdenhil, and W. Wagner, “Comparison of long short-term memory networks and random forest for sentinel-1 time series based large scale crop classification,” Remote Sens (Basel), vol. 13, no. 24, p. 5000, 2021.

[6] A. Kamilaris and F. X. Prenafeta-Boldú, “A review of the use of convolutional neural networks in agriculture,” J Agric Sci, vol. 156, no. 3, pp. 312–322, 2018, doi: DOI: 10.1017/S0021859618000436.

[7] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[8] L. Breiman, “Random forests,” Mach Learn, vol. 45, pp. 5–32, 2001.