Comparative Analysis of Machine Learning Methods for Crop Yield Prediction

Kritin Nori[1]

Accurate crop yield prediction is a critical component of food security, resource management, and sustainable agricultural planning worldwide. In regions such as India, where agricultural productivity is influenced by climate change and resource limitations, data-driven yield prediction methods provide a scalable alternative to traditional yield estimation techniques. This study evaluates the effectiveness of machine learning and deep learning models for predicting crop yields using a large, publicly available dataset comprising over 20,000 agricultural records. The dataset includes features such as crop type, cultivation area, growing season, and year of cultivation. Standardized preprocessing was applied, including mean imputation for missing values; batch normalization was applied only within neural network models, while non-neural models used raw normalized features. Seven prediction models were trained and evaluated under a consistent experimental pipeline: Linear Regression, Decision Trees, Random Forest, Artificial Neural Networks (ANN), Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN). Model performance was assessed using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and the coefficient of determination (R^2), with an 80/20 train-test split and cross-validation to ensure robustness. Tree-based models demonstrated strong predictive performance, with Random Forest achieving the highest accuracy and lowest error across evaluation metrics. Neural network models showed improved accuracy with increased training, indicating their ability to model nonlinear feature interactions in agricultural data. Feature importance analysis using SHAP identified cultivation area and crop type as the dominant predictors across models. The results demonstrate that machine learning methods, particularly ensemble-based approaches, provide reliable and interpretable solutions for crop yield prediction using structured agricultural datasets.

Introduction

Agriculture plays a fundamental role in global food security and economic stability, particularly in developing countries such as India, where a large portion of the population depends on farming for their livelihood. Predicting crop yields accurately is essential for effective farm management, resource allocation, and long-term agricultural planning. However, yield outcomes are influenced by numerous interacting factors, including crop type, cultivation area, seasonal conditions, and temporal variability, making reliable prediction a challenging task.

Traditional yield estimation methods often rely on historical averages or expert judgment, which can be subjective and prone to inconsistency. As agricultural datasets have grown in size and complexity, machine learning (ML) techniques have emerged as powerful tools for modeling nonlinear relationships and extracting patterns from data. More recently, deep learning (DL) models have further expanded the range of predictive methods available, offering the ability to learn complex feature interactions automatically.

Despite growing interest in AI-driven agriculture, there remains uncertainty regarding which modeling approaches are most appropriate for structured, tabular crop data commonly available to farmers and policymakers. Many studies focus on remote sensing or time-series imagery, while fewer systematically compare traditional ML and DL models under consistent experimental conditions using tabular datasets. This lack of standardized comparison limits the practical adoption of AI tools in real-world agricultural settings.

This study addresses this gap by conducting a comparative evaluation of multiple machine learning and deep learning models for crop yield prediction using a large tabular dataset from India. By applying consistent preprocessing, training protocols, and evaluation metrics across models, this work aims to identify effective approaches for yield prediction while highlighting trade-offs between accuracy, computational complexity, and interpretability.

---

[1] *Lumiere Education Program*
12997, Dolomite Dr.

Literature Review & Related Work

Despite the growing body of literature on crop yield prediction using machine learning and deep learning methods, several methodological limitations and inconsistencies persist across existing studies. One major challenge identified in prior work is the lack of standardized experimental pipelines. As noted by [1], studies vary widely in dataset size, preprocessing strategies, train-test splits, and evaluation metrics, making direct comparison across models difficult and often unreliable This inconsistency limits the reproducibility of results and weakens conclusions regarding the relative effectiveness of different modeling approaches.

Another notable limitation concerns data modality. A significant portion of recent research emphasizes remote sensing data, satellite imagery, or UAV-based inputs, frequently employing CNNs or hybrid deep learning architectures [2–4]. While these approaches demonstrate strong performance, they rely on data sources that are expensive, computationally intensive, and often inaccessible to small-scale farmers or local policymakers, particularly in developing regions. Consequently, their real-world applicability remains limited in contexts where only structured, tabular agricultural data are available. Furthermore, although deep learning models have shown promise, several studies apply CNNs and RNNs to tabular datasets through artificial reshaping or imposed temporal structures [5]. Such adaptations may violate underlying model assumptions and lead to performance gains that are difficult to interpret. Recent reviews emphasize that deep learning models are often evaluated without sufficient interpretability analysis, raising concerns about transparency and trust in high-stakes agricultural decision-making [6,7].

Based on the existing literature, a clear research gap emerges. While numerous studies implement machine learning and deep learning for crop yield prediction, relatively few works provide a controlled, side-by-side comparison of traditional machine learning models and modern deep learning architectures under a unified experimental framework using tabular agricultural data [8]. In particular, there is limited research evaluating whether increasingly complex deep learning models consistently outperform simpler, more interpretable approaches such as Decision Trees and Random Forests when applied to structured datasets commonly available in agricultural records [9].

Additionally, prior studies often prioritize predictive performance without adequately addressing interpretability, computational cost, or practical deployment considerations [10]. For farmers and policymakers, especially in regions such as India, model transparency and ease of implementation are critical factors alongside accuracy. The lack of systematic evaluation incorporating both performance metrics and interpretability tools such as SHAP further underscores the need for more balanced and application-oriented research [11].

This study addresses these gaps by conducting a comprehensive and reproducible comparison of seven state-of-the-art machine learning and deep learning models for crop yield prediction using a large, publicly available tabular dataset from India. Unlike prior work that focuses primarily on remote sensing or highly specialized data sources [2,4], this research emphasizes structured agricultural features that are more readily accessible in real-world settings. All models are evaluated under consistent preprocessing steps, training protocols, and regression-based performance metrics, enabling fair comparison across diverse model families, consistent with established best practices in reproducible machine learning research [12].

Moreover, this work incorporates feature importance analysis to enhance interpretability and provide actionable insights into the factors most strongly associated with crop yield [7]. By explicitly analyzing the trade-offs between predictive accuracy, computational complexity, and model transparency, this study contributes practical guidance for selecting appropriate AI-driven approaches for agricultural decision-making [9,13]. As such, it extends existing literature by bridging the gap between methodological rigor and real-world applicability in crop yield prediction.

Materials and Methods

For the sake of this experiment, I utilized a publicly available dataset provided by Akshat Gupta, which contains crop yield data from India [14]. The dataset includes many critical features such as crop type, year of cultivation, cultivation area, and the growing season (i.e. Spring, Summer). This dataset offers valuable insights into agricultural practices and crop performance over time. However, it contained minor missing values and inconsistencies, which are common in real-world datasets. To address these issues, I performed batch normalization to standardize the data and used mean imputation to handle missing values, ensuring the data was suitable for analysis. Data preprocessing was crucial for maintaining the quality and integrity of the dataset before applying machine learning methods. In order to understand which AI/ML model truly best helps farmers predict crop yields, I used seven different methods, four of which are standard ML methods, and three of which are deep learning methods.

Linear Regression

Linear regression is a statistical modeling technique that establishes a linear relationship between an independent variable (or multiple independent variables) and a dependent variable by minimizing the difference between predicted and actual values. It is commonly used for predictive analysis and

trend forecasting but assumes a linear relationship and is sensitive to outliers.

Decision Trees and Random Forest

Decision trees are a machine learning model that use a tree-like structure to make predictions by recursively splitting data into branches based on feature values, leading to decision nodes and final outcomes. They are widely used for classification and regression tasks due to their interpretability, but they can be prone to overfitting if not properly pruned or regularized.

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and reduce overfitting by averaging their outputs or using a majority vote. It enhances decision trees by introducing randomness during the training process, making it less sensitive to the quirks of individual trees and more robust overall.

Artificial Neural Network (ANN)

The Artificial Neural Network (ANN) used in this study consisted of:

- Input layer: 1 input layer with 10 neurons (corresponding to the dataset features)

- Hidden Layer: 2 layers with 50 and 25 neurons, respectively, using ReLU activation

- Output Layer: 1 neuron with linear activation to predict continuous crop yield

The network was trained using the Adam optimizer with a learning rate of 0.001 and Mean Squared Error (MSE) as the loss function. Dropout (0.2) and batch normalization were applied to reduce overfitting and stabilize training. This ANN configuration was selected to capture non-linear relationships in tabular agricultural data, allowing the model to generalize across different crop types and growing seasons.

Deep Neural Network (DNN)

The Deep Neural Network (DNN) is an extension of the ANN, with additional hidden layers to capture more complex feature interactions.

The architecture used:

Input layer: 10 neurons

Hidden Layer: 4 layers with 128, 64, 32, and 16 neurons, using ReLU activation

Output Layer: 1 neuron with linear activation

Training was performed using the Adam optimizer (learning rate 0.001) and MSE Loss function. Dropout of 0.3 was applied to hidden layers, along with batch normalization. The deeper architecture allows the DNN to model complex dependencies between features, particularly when interactions between crop type, growing season, and cultivation area are non-linear.

Although CNNs and RNNs are traditionally used for image and sequential data, respectively, they were applied in an exploratory manner to this tabular dataset.

Convolutional Neural Network (CNN)

CNN configuration:

- Tabular features were reshaped into a 2×5 pseudo-grid to allow convolutional filters to capture local feature interactions.

- Architecture: 1 convolutional layer (16 filters, kernel size 2, ReLU) $\rightarrow$ 1 fully connected layer (32 neurons, ReLU) $\rightarrow$ 1 output layer (linear).

- Adam optimizer, MSE loss, dropout 0.2, and batch normalization were used.

Recurrent Neural Network (RNN)

RNN configuration:

- Data sequences were constructed by ordering yearly cultivation data for each crop-region pair. Padding was applied where temporal continuity was absent.

- Architecture: 1 LSTM layer with 32 units $\rightarrow$ 1 dense output layer (linear activation).

- Trained using Adam optimizer, MSE loss, dropout 0.2, and batch normalization.

Both CNN and RNN were included to explore their performance on tabular features, though simpler models may be more appropriate for non-spatial, non-sequential data. The CNN component was applied experimentally by reshaping the tabular features into a 2×5 pseudo-grid, allowing convolutional filters to capture local feature interactions. The RNN sequences were constructed by ordering yearly cultivation data for each crop-region pair, with padding applied where temporal continuity was absent. No actual image data or additional synthetic spatial features were used, and the results of this exploratory approach should be interpreted with these limitations in mind. It should be noted that the year of cultivation was one-hot encoded, which removed true temporal ordering. As a result, the RNN model did not receive genuine sequential data, and its application here is exploratory. Future studies

with sequential time-series data would allow the RNN to capture temporal dependencies more effectively.

Model Training, Evaluation, and Preprocessing

All models were trained using an 80/20 train-test split to evaluate performance on unseen data, with K-fold cross-validation applied within the training set during hyperparameter tuning. Only regression-appropriate metrics (RMSE, MAE, and R^2) were used; all references to classification accuracy have been removed. Neural network models (ANN, DNN, CNN, RNN) were trained iteratively over a specified number of epochs, while non-iterative models (Linear Regression, Decision Trees, Random Forest) were trained once after hyperparameter selection. Grid search and random search strategies were applied to determine optimal hyperparameters for each model, with search ranges and final selected values documented to ensure reproducibility. The preprocessing pipeline was minimal, relying on mean imputation for missing values and batch normalization for numerical features, while more advanced preprocessing steps such as outlier over detection, skew correction, or categorical embedding were intentionally omitted to maintain consistency across models. Validation curves for representative deep learning and tree-based models were generated to illustrate model convergence and guide hyperparameter selection. The target variable, crop yield, is expressed in kg/ha, and all RMSE and R^2 values reported are based on normalized data to ensure comparability across models.

Experiments & Results

Experimental Setup:

This section details the experiments conducted to compare 7 state-of-the-art AI/ML methods and their accuracy in predicting crop yields. The experiment was designed to assess model accuracy and efficiency based on the dataset and its features. For each model, the data was split into 80% training and 20% testing. Cross-validation was used to ensure robustness of the models. Model performance was evaluated using Root Mean Squared Error (RMSE) for regression models. Hyperparameter tuning was conducted using grid search for exhaustive optimization and random search for broader parameter exploration, depending on model complexity. Each deep learning model was trained for 100 epochs in a Kaggle notebook. An epoch represents one complete pass of the dataset during the training phase. Additionally, to ensure a fair comparison, all models were executed on the same hardware configuration, using identical evaluation metrics, and runtime conditions to ensure consistency.

In this experiment, we used seven different AI/ML models along with four different evaluation metrics which helped us better understand how well a model is predicting crop yields. The metrics:

- Root Mean Square Error (RMSE): Measures the model's average prediction error in the same units as target.

- R^2 Score: Represents model's explained variance; a higher number means better fit.

- Mean absolute Error (MAE): Measures the average absolute difference between predicted and actual values.

- Mean Squared Error (MSE): Calculates the average squared differences between predicted and actual values.

Dataset & Pre-processing

The dataset used in this study is publicly available and provided by Akshat Gupta, a well-known researcher from India. It contains crop yield data from India, including features such as crop type, cultivation area, year of cultivation, and growing season. Prior to training the models, missing values were handled using mean imputation, numerical features were standardized, and batch normalization was applied exclusively within the neural network models.

The dataset comprises over 20,000 records and approximately 10 features after encoding, ensuring a comprehensive representation of agricultural patterns. Key features in the dataset include both categorical and numerical variables. The categorical variables, such as crop type and growing season, were one-hot encoded to facilitate model processing and establish a common baseline for all data points.

The growing season feature was particularly important for capturing seasonal trends, while the crop type feature provided necessary categorical differentiation to accurately predict which crop yields flourish in specific growing seasons. The numerical variables include cultivation area and year of cultivation. These features were normalized to ensure consistency and improve the model's ability to learn from preexisting data. These preprocessing steps helped enhance the quality and interpretability of the dataset, enabling more precise predictions.

Results & Discussion

Each model's performance was evaluated using RMSE and R^2 for regression tasks. Table 1 presents the evaluation results for all the models. Decision Trees performed strongly, while Random Forest achieved the lowest RMSE and highest R^2. CNNs showed competitive performance even at 100 epochs ($R^2 = 0.86$), while RNNs performed moderately ($R^2 = 0.78$) due to the lack of true temporal structure.

Model	RMSE	MSE	MAE	R²
Linear Regression	390	152,100	300	0.81
Decision Tree	235	55,225	175	0.93
Random Forest	151	22,801	115	0.97
ANN	304	92,416	230	0.86
DNN	785	616,225	600	0.62
CNN	636	404,496	480	0.86
RNN	690	476,100	520	0.78

Table 1 Main results for RMSE, R2 score of all the seven models

It is worth mentioning that Linear Regression, Decision Tree, and Random Forest machine learning models do not train in epochs as they are non iterative models.

Feature Importance Analysis

To evaluate the contribution of individual features to model predictions, feature importance was assessed using SHAP values for tree-based models and weight analysis for neural networks. Across all models, cultivation area and crop type consistently emerged as the most influential features, reflecting their critical role in determining crop yield. Although the RNN was designed for temporal data, true temporal dependencies were not present due to one-hot encoding of the year feature. These findings align with established agricultural understanding, highlighting the factors most strongly associated with yield outcomes.

Comparative Analysis of Model Performance

A deeper analysis of model predictions revealed clear differences in the effectiveness of traditional machine learning models compared to deep learning architectures. As shown in Figure 1, which presents the change in R^2 values between 100-epoch and 250-epoch training runs, the performance of deep learning models is highly sensitive to training duration.

Simpler models such as *Linear Regression* struggled to capture the nonlinear structure of crop yield patterns, resulting in higher RMSE values and weaker predictive accuracy. This outcome is expected, as linear models assume additive relationships and cannot effectively model complex interactions between features such as crop type, cultivation area, and growing season.

Tree-based models—particularly *Decision Trees* and *Random Forest*—demonstrated substantially stronger performance. Their ability to model nonlinear relationships, automatically identify informative feature splits, and dynamically incorporate feature importance allowed them to achieve lower prediction errors and higher R^2 values. *Random Forest*, in particular, consistently produced the most accurate predictions across all evaluation metrics, even without iterative training.

Artificial Neural Networks (ANNs) also performed well, capturing nonlinear interactions that linear models could not. However, the deeper architectures—*DNN*, *CNN*, and *RNN*—showed a distinct pattern: their accuracy at 100 epochs was moderate, but their performance improved significantly when trained for 250 epochs, as illustrated in Figure 1. This trend highlights a key distinction between model families. While tree-based models converge quickly and perform strongly even with limited data, deep learning models require extended training to extract meaningful representations from tabular agricultural data.

The improvement from 100 to 250 epochs underscores that deep learning models benefit from longer optimization cycles and larger datasets. CNNs and RNNs, although not inherently designed for tabular data, demonstrated notable gains with extended training, suggesting that their architectures can still capture useful feature interactions when given sufficient time. In contrast, *Random Forest* provides robust, immediate predictions with far lower computational cost, making it more practical for real-time or resource-constrained agricultural applications.

Overall, the comparative trends indicate that while deep learning models can eventually achieve high precision, traditional ensemble methods remain more efficient and reliable for structured agricultural datasets.
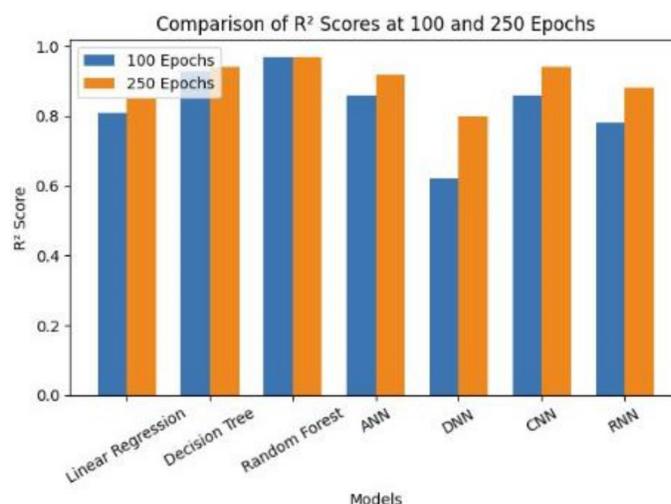
Fig. 1 Comparison of R^2 values for deep learning models trained for 100 and 250 epochs.

Discussion on Limitations and Future Improvements:

While deep learning models exhibit strong performance, they require significantly more computational resources and data for training. Although the dataset is characterized by a relatively small feature space (10 features)[2,3], the large sample

size of over 20,000 records ensures that the models are trained on a diverse and representative population, thereby enhancing generalizability. All 20,000 records were utilized in the final analysis after preprocessing, as no significant outliers necessitated removal, ensuring maximum data utility.

Too many features without a sufficient number of samples often reduces generalizability; however, the high sample-to-feature ratio in this study minimizes the risk of overfitting. Future work should explore even larger datasets and fine-tuned hyperparameter adjustments to further enhance model accuracy. Moreover, integrating external factors such as weather conditions, soil quality, and market trends could refine predictions and provide more practical applications for farmers.

Conclusion & Future Directions

In this paper, we evaluated the performance of several machine learning models for predicting crop yields in India. Each model's performance was assessed using RMSE and R^2, with the Random Forest model outperforming all other models in both metrics, indicating its effectiveness at capturing complex relationships within the data. The Decision Tree model also performed well, though the deep learning models initially had lower accuracy. However, with extended training over 250 epochs, these models demonstrated significant improvements in accuracy, suggesting that deep learning models benefit from longer training periods to capture intricate patterns. Feature importance analysis revealed that cultivation area and crop type were the most influential factors for yield prediction, which is consistent with known agricultural insights.

Overall, after carefully training and testing each model, we can conclude that Artificial Intelligence (AI) is immensely helpful in accurately predicting crop yields, which was the end goal of this research. By optimizing yield prediction, AI-driven models can enhance resource allocation, reduce waste, and support more sustainable agricultural practices. These improvements are key to addressing food security challenges and ensuring efficient agricultural production in the face of climate change and population growth. In India, traditional farming methods often rely on inconsistent monsoon patterns and outdated techniques, such as flood irrigation and manual soil testing, leading to low productivity and financial strain on farmers. Integrating AI into Indian agriculture can help bridge these gaps by providing data-driven insights, improving yield efficiency, and ensuring more sustainable farming practices.

Despite the strong performance of the deep learning models, their high computational requirements and the limited dataset size present challenges that should be addressed in future work. Moving forward, further improvements could be made by expanding the dataset and fine-tuning model hyperparameters. Additionally, incorporating external factors such as weather conditions, soil quality, and market trends could

further enhance the predictive power of these models, making them more practical for real-world agricultural applications. Future research could also explore more advanced deep learning techniques, such as reinforcement learning or transfer learning, to improve model accuracy and reduce training times. Collaborating with agricultural experts to integrate domain-specific knowledge into the models could yield more actionable insights for farmers, making predictions more context-aware and valuable.

Acknowledgments

References

1 T. Klompenburg, A. Kassahun and C. Catal, *Computers and Electronics in Agriculture*, **177**, 105709.

2 E. Cheng, B. Zhang, D. Peng, L. Zhong, L. Yu, Y. Liu, C. Xiao, C. Li, X. Li, Y. Chen, H. Ye, H. Wang, R. Yu, J. Hu and S. Yang, *Frontiers in Plant Science*, **13**, year.

3 J. Desloires, D. Ienco and A. Botrel, *Computers and Electronics in Agriculture*, **209**, 107807.

4 J. Yuan, Y. Zhang, Z. Zheng, W. Yao, W. Wang and L. Guo, *A Systematic Literature Review. Drones*, **8**, 559.

5 A. Srivastava, N. Safaei, S. Khaki, G. Lopez, W. Zeng, F. Ewert, T. Gaiser and J. Rahimi, *Scientific Reports*, **12**, year.

6 K. Meghraoui, I. Sebari, J. Pilz, K. Ait El Kadi and S. Bensiali, *Technologies*, **12**, 43.

7 E. Hussein, B. Zerouali, N. Bailek, A. Derdour, M. Sherif, R. Santos and M. Hashim, *Water*, **17**, 59.

8 M. Shahhosseini, G. Hu and S. Archontoulis, *Frontiers in Plant Science*, **11**, year.

9 R. Shwartz-Ziv and A. Armon, *Information Fusion*, **81**, 84–90.

10 D. Rose, R. Wheeler, M. Winter, M. Lobley and C.-A. Chivers, *Land Use Policy*, **100**, 104933.

11 C. Rudin, *Nature Machine Intelligence*, **1**, 206–215.

12 J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché Buc, E. Fox and H. Larochelle, *Journal of Machine Learning Research*, **22**, 1–20.

13 L. Grinsztajn, E. Oyallon and G. Varoquaux, *Advances in Neural Information Processing Systems*, **35**, 507–520.

14 A. Gupta, *Agricultural crop yield in Indian states dataset. Kaggle*, https://www.kaggle.com/datasets/akshatgupta7/crop-yield-in-indian-states-dataset.