

Enhanced Crop Yield Prediction using Hybrid Deep Learning Models

Sagana C¹, Thangatamilan M², Sangeetha M³, Tharun Kumar S⁴, Suhas C⁵ and Vishal M⁶

^{1,2,4,6}Kongu engineering college, Perundurai, Tamil Nadu, India

Email: sagana.c@gmail.com, thangatamil.m@gmail.com

³Dr. N. G. P Institute of Technology, Coimbatore, India

Email: sangeethcse@gmail.com

Abstract—India's economy depends primarily on agriculture. Crop yield offers significant potential for food production worldwide. It enables farmers and policymakers to make informed decisions to improve food security and maximize resource allocation. The yield of crops depends heavily on multiple factors such as climate conditions, rainfall, soil nutrients, agricultural practices, etc. The climate has a major impact on the crop's potential production. A hybrid LSTM model has been developed for crop yield prediction and trained on a diverse dataset comprising historical yield records and meteorological data. Experimental results show that hybrid LSTM models outperform the standalone models, such as CNN, GRU, and LSTM, in crop yield prediction, and the model performance is measured using MSE, RMSE, MAE, and R2. The Hybrid LSTM Model has an MSE of 0.0015, RMSE of 0.03, MAE of 0.007, and R-squared value of 0.86.

Keywords— Agriculture, crop yield prediction, Deep learning, long short-term memory, CNN, GRU, Hybrid LSTM.

I. INTRODUCTION

Agriculture serves as the foundation of India's economy, and it is the main source of employment for over half of the population and supports millions of people across the nation. Crops grown by farmers today are based on past knowledge and often depend on traditional methodologies without incorporating data-driven insights. Such yields have been abundant due to the absence of an accurate demand assessment for cultivating specific crops under the conventional approach. [1]. These challenges are further worsened to feed a rapidly increasing population in the face of decreasing natural resources and the looming spectra of climate change. With the rising population growth, there will be a demand for food globally, which will cause food insecurity in densely populated areas. To address these issues, decision-makers in agriculture and farmers need essential tools to determine the most suitable crops for cultivation based on past yield patterns and environmental conditions [6]. Predicting crop production is essential to modern agriculture because it ensures environmental sustainability, economic stability, and food security [2]. By accurately forecasting crop yields, stakeholders can anticipate potential shortages, enabling proactive measures to be taken to ensure adequate food supply for communities and nations. Moreover, it allows farmers to plan their resources effectively, minimizing wastage and optimizing production to ensure stable incomes. Precise forecasts can assist in maximizing farming methods, reducing losses, and guaranteeing food security despite population growth and climate change.

Artificial Intelligence (AI) techniques have revolutionized estimating crop yields, considering several agricultural issues examined by Javed, Md Abu [9]. Deep learning has emerged as a transformative tool in crop yield prediction, offering unparalleled accuracy and efficiency in analyzing complex agricultural data and addressing the complexities posed by crop diversity, geographical variations, and cultivation areas. To identify hidden patterns and correlations, deep learning models can process vast datasets, including satellite imagery, weather patterns, soil characteristics, and historical yield records. DL techniques like Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and hybrid models enable real-time insights, supporting sustainable farming practices and optimizing agricultural productivity. By leveraging these advanced methods, farmers and policymakers can make data-driven decisions to improve resource allocation and enhance food security.

In this paper, we present a hybrid LSTM model for crop yield prediction that combines the strengths of convolutional neural networks and other machine learning techniques, Neural networks (CNNs), and long short-term memory. Several factors are considered to forecast the crop yield, including crop name, season, district, area, production, and yield, to enhance prediction accuracy. We compare our hybrid model's performance with isolated models, e.g., CNN, GRU, and LSTM, in terms of evaluation metrics like Mean Absolute Error (MAE), Mean Squared Error (RMSE), Root Mean Squared Error (RMSE), and R-squared error (R²).

II. LITERATURE REVIEW

V. Selam et al. [3] implemented the significance of yield forecasts for farmers to minimize losses and maximize crop prices. The factors influencing crop output are environmental data, such as Area under Cultivation, Food Price Index, and Annual Rainfall. Regression analysis is a multivariate method used to examine various environmental factors and how they affect crop yield. Pant, Janmejaya, et al. [2] Various machine learning algorithms are employed to forecast crop yield in India. [2]. The features such as area, Variety of crop and production year, yield, average rainfall, Number of pesticides used, and average temperature have been used. The forecast for four major crops, including potatoes, rice, wheat, and maize. The decision tree regressor attains the highest level of accuracy in forecasting crop yield.

Iniyar et al [1] investigate many ways to forecast crop yields using various soil and environmental factors. Farmers can utilize machine learning algorithms to determine the most suitable crop to cultivate while simultaneously enhancing their harvest. Farmers can gain advantages from yield estimation as it enables them to reduce crop loss and secure the highest possible prices for their produce. To overcome the difficulties in accurately predicting agricultural production, Patrick Helber et al. [3] suggested remote-sensed information in conjunction with models for machine learning, like Support Vector Machines (SVM) and decision trees. To attain a global yield model, this method does not require crop-specific models. Using information from remote sensing and machine learning models like Support Vector Machines (SVM) and decision trees, they have an R² value of 0.77 and an RRMSE of roughly 15% at the field level. To attain a global yield model, crop-specific models are used. They have an error rate (RRMSE) of roughly 15% and an R² value of 0.77 at the field level. indicating their efficacy across a range of crop kinds. This development has the potential to maximize crop yields by offering accurate yield estimates and projections.

Alexandros et al. [4] proposed Hybrid Deep Learning-Based Models for Crop Yield Prediction. CNN-XGBoost, CNN-Recurrent Neural Networks (RNN), CNN-Deep Neural Networks (DNN), CNN-Long Short-Term Memory (LSTM), and the XGBoost machine learning (ML) algorithm were among the deep learning-based models executed. The hybrid CNN-DNN model outperforms the other models, depending on the results.

Saeed Khaki et al., [5] the authors sought to correctly estimate crop yield with Deep neural network (DNN), Lasso, and Shallow neural network. They worked with files with genotype and yield data required maize hybrids from various regions and years of the 2018 Syngenta Crop Challenge. They used a deep neural network (DNN) approach and expanded modeling and solution techniques. to develop a predictive model. El Sayed M. El Kenawy et al [7] predicted Potato Crop Yield using Machine learning and Deep Learning. Various predictive models like K-nearest neighbors (KNN), gradient boosting, XGBoost, and multilayer perceptrons are utilized as well as long short-term memory networks (LSTM), graph neural networks (GNNs), and gated recurrent units (GRUs) to predict the yield.

A deep learning approach for crop production prediction supported by genetic algorithms (GA) is proposed by Bi Luning et al [8]. Two stages of such a model involve a worldwide search stage with GA for finding perfect initial neural network weights and a local search stage with perturbative randomness for resolving vanishing gradient and nearby optima concerns. Concerning gradient-based approaches, increased prediction accuracy and a higher velocity of convergence of the proposed model have been established. C. Sagana et al [11]

recommended using machine learning algorithms to increase crop yield and productivity by considering different parameters like weather, soil, and environmental characteristics.

III. METHODOLOGY

A. Data Collection and Pre-processing

The heart of our Crop Prediction System is the Data Collection and Preprocessing module, which starts with the monotonous acquisition of a massive dataset of 8000 rows and 7 columns before pre-processing. At this stage, data must be gathered from reliable sources, data correctness must be checked, and preprocessing is done. The data is structured, sanitized properly, and checked for outliers to ensure its integrity.

B. Hybrid LSTM model

The data is placed into a hybrid LSTM model depicted in Figure 1 after preprocessing. For effective crop yield prediction, this hybrid model involves various components, such as batch normalization, LSTM layers, dropout layers, and embedding layers. The layer of pre-processing further refines the data for the trained model, which may involve tasks such as embedding categorical features (Crop Type, District, Season) into dense vector representations. Helps improve model efficiency by reducing dimensionality and capturing relationships between different categories.

Embedding layer: Categorical features (Crop Type, District, Season) are embedded using an embedding layer, which transforms them into dense numerical representations for better processing. Embeddings help capture relationships between different categories in a lower-dimensional space. These embeddings are then concatenated with numerical inputs such as year, area, and production.

LSTM layer: LSTM networks are particularly adept at gathering Long-term sequential dependencies. It is aware of the temporal relationship of the dataset, maintains an internal memory state and enables the model to be able to learn patterns between past and current inputs to make better predictions, thus it's best suited for time-series data.

Dropout Layer: A regularization method that randomly eliminates neurons during the training process to avoid overfitting and guarantees that the model performs well on new, unseen data.

Batch Normalization: A technique used to normalize activations within the network, improving stability and convergence speed. Helps reduce overfitting and ensures that gradients flow smoothly, making training more efficient.

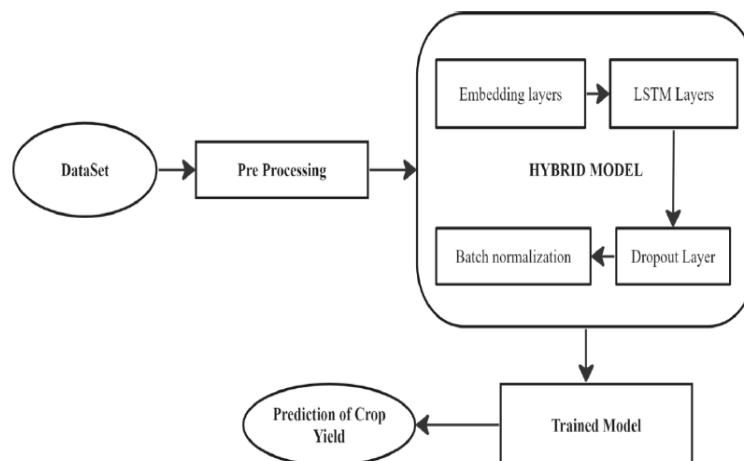


Figure 1. Flow diagram of the proposed system

C. Deep Learning Models

For the deployment of the crop yield prediction model, higher-level deep learning models such as Convolutional Neural Networks (CNNs), Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), and hybrid LSTM are recommended. CNNs learn the spatial features and patterns from multidimensional data. GRU networks offer the same level of performance as LSTMs but are computationally more efficient as they have fewer architectural

components, hence are best suited for applications where the computing resources are limited. LSTM networks are inherently suitable to capture temporal factors like seasonal variations and other dynamics play a role in shaping the patterns observed in the data and production history, and can learn long-term dependencies in sequential data, predicting crop yield. The hybrid LSTM Model integrates LSTM layers with traditional dense layers, providing a potent framework for modeling both temporal dependency and nonlinear relationships in the data. By integrating LSTM's memory-enhancing capability and representational power of dense layers, the hybrid LSTM model can successfully learn intricate patterns from various sources of information, including yearly information and categorical variables like crop, season, and district. The integration promises improved predictive power and strength, and the Hybrid LSTM model is an effective and strong tool for crop yield prediction in difficult agricultural environments.

IV. EXPERIMENT ANALYSIS

In evaluating the results, a set of experiments is conducted to evaluate the performance of the existing models and the proposed hybrid LSTM model integrated with embedding layers, LSTM layers, dropout, and batch normalization. This enhanced model improves predictive accuracy and minimizes loss values compared to the standard baseline models. The Hybrid LSTM Model has an MSE of 0.0015, RMSE of 0.03, MAE of 0.007, and R-squared value of 0.86.

A. Dataset description

For implementing the Crop Yield Prediction System, relevant data are gathered from the official government website which contains crop name, area, production, and yield data from different districts of TamilNadu. These datasets undergo preprocessing to handle missing values, standardize units, encode categorical variables "Crop," "District," and "Area," normalize numerical features, and Min-Max scaling to standardize the target variable "Yield" to ensure the quality and consistency of the dataset for subsequent analysis.

B. Evaluation Matrices

To compare the performance of the deep learning models, metrics such as Mean Squared Error (MSE), R-squared (R²), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) are considered in evaluating the proposed work.

MSE (Mean Squared Error)

The average size of prediction errors can be measured by the mean squared error (MSE), a measure of the mean squared difference between expected and actual values. Larger errors are treated more severely.

$$MSE = \frac{1}{N} \sum_{i=1}^N (\text{predicted}_i - \text{actual}_i)^2$$

MAE (Mean Absolute Error)

The average absolute difference between the expected and actual values is measured by the Mean Absolute Error or MAE. Without considering the account of their direction, it gives information on the typical size of prediction errors.

$$MAE = \frac{1}{N} \sum_{i=1}^N |\text{predicted}_i - \text{actual}_i|$$

R-Squared

R-squared is a statistic that gives the proportion of variation in the dependent variable explained by the independent variables in a regression model. It is a measure of the goodness of fit of the model.

$$R^2 = 1 - \frac{\sum (\text{actual}_i - \text{predicted}_i)^2}{\sum ((\text{actual}_i - \text{mean of actual})^2)}$$

Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is a statistical indicator to measure the performance of a regression model. It is the square root of the mean of the squared prediction errors, i.e., the differences between predicted and actual values. RMSE gives a measure of the spread or dispersion of prediction errors. It is especially useful as it punishes larger errors more than smaller ones, so it is sensitive to outliers.

$$RMSE = \sqrt{\sum_{i=1}^N \frac{(predicted_i - actual_i)^2}{Nx}}$$

C. Results

The mean squared error for four deep learning models LSTM, CNN, GRU, and Hybrid LSTM for crop yield prediction is shown in Figure 2. A lower mean squared error (MSE) is generally indicative of a better-performing model. Hybrid LSTM appears to be the best model as it integrates LSTM layers with dense layers for enhanced prediction accuracy. This model has the lowest MSE Training value of 0.01, and a test value of 0.015 among all, indicating superior performance and excellent generalization.

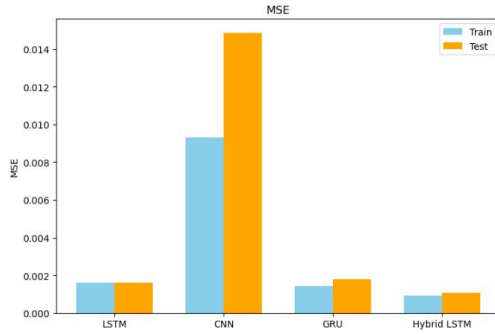


Figure 2. MSE comparison

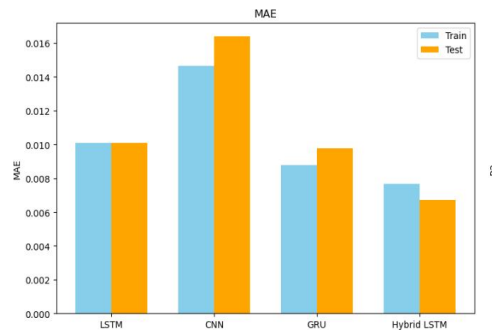


Figure 4. MAE Comparison

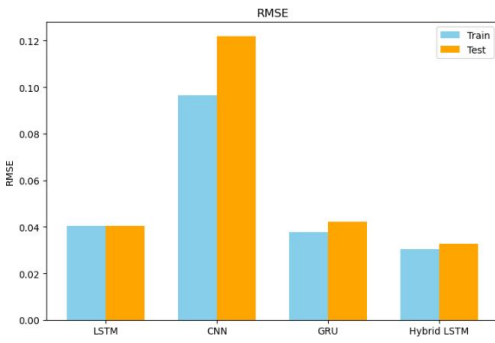


Figure 3. RMSE Comparison

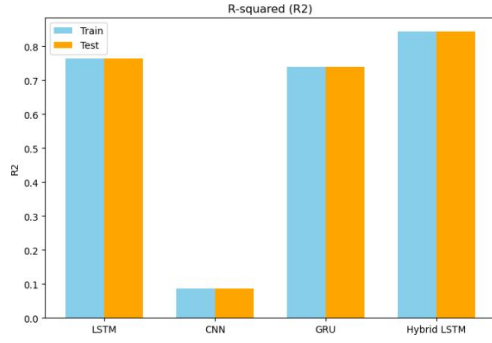


Figure 5. R squared error comparison

Figure 3 presents the Root Mean Squared Error (RMSE) for four deep learning models LSTM, CNN, GRU, and Hybrid LSTM in both training and testing phases. The hybrid LSTM model has the lowest RMSE Training value of 0.03, a test value of 0.035 among all, indicating superior performance and excellent generalization.

Figure 4 compares four models—Hybrid LSTM, GRU, CNN, and LSTM—in predicting crop yields based on MAE for both training and testing datasets. It shows that the hybrid LSTM is performing higher than the other models, which has the lowest MAE Training value of 0.0075, test value of 0.007 among all other models.

Figure 5 compares four models—hybrid LSTM, GRU, CNN, and LSTM—in predicting crop yields based on R-squared for both training and testing datasets. As we know, a higher value means that the model is better at capturing and explaining the variation in the data. It shows that hybrid LSTM is way more efficient than LSTM and CNN in R-squared comparison.

TABLE I. ERROR VALUE FOR LSTM, CNN, GRU, AND HYBRID LSTM MODEL

Model	MSE		RMSE		MAE		R ²	
	Train phase	Testing phase	Train phase	Testing phase	Train phase	Testing phase	Train phase	Testing phase
LSTM	0.002	0.002	0.04	0.04	0.01	0.01	0.75	0.76
CNN	0.009	0.009	0.10	0.12	0.015	0.017	0.10	0.11
GRU	0.002	0.002	0.035	0.045	0.09	0.10	0.73	0.74
Hybrid LSTM	0.0015	0.0015	0.03	0.035	0.007	.0065	0.86	0.87

Table 1 shows the error value of the training phase and testing phase of LSTM, CNN, GRU, and hybrid LSTM models. For all the MSE, RMSE, MAE, and R2, the Hybrid LSTM model shows the lowest MSE (0.0015), lowest RMSE (0.03-0.035), and lowest MAE (0.007-0.0065), indicating minimal prediction errors. The R² score of 0.86 in training and 0.87 in testing suggests it explains the highest variance in the dataset, making it the most effective model for this task.

V. CONCLUSION

A robust Crop Yield Prediction System is developed to tailor the agricultural landscape of Tamil Nadu by implementing cutting-edge deep learning methods, including embedding layers, LSTM networks, dropout, and batch normalization. We have achieved notable improvements in predictive accuracy and model performance. The inclusion of embedding layers facilitated the capture of intricate data relationships, while LSTM networks effectively modelled temporal dependencies, resulting in more accurate yield predictions. Furthermore, incorporating dropout and batch normalization techniques ensured model stability and reduced overfitting, enhancing overall prediction reliability. Compared with LSTM, CNN, and GRU, Hybrid LSTM generates the lowest errors and the highest R² score, making it the most accurate and reliable. The model may effectively acquire and represent complicated spatial and temporal data thanks to this architecture. relationships in agricultural data, making it well-suited for accurate yield prediction.

REFERENCES

- [1] Iniyan, S., V. Akhil Varma, and Ch Teja Naidu. "Crop yield prediction using machine learning techniques." *Advances in Engineering Software* 175 (2023): 103326.
- [2] Pant, Janmejay, et al. "Analysis of agricultural crop yield prediction using statistical techniques of machine learning." *Materials Today: Proceedings* 46 (2021): 10922-10926.
- [3] Helber, P. et al. (Year). Crop Yield Prediction: An Operational Approach To Crop Yield Modeling On Field And Subfield Level With Machine Learning Models.
- [4] Alexandros Oikonomidis, Cagatay Catal & Ayalew Kassahun (2022) Hybrid Deep Learning-based Models for Crop Yield Prediction, *Applied Artificial Intelligence*, 36:1, 2031822, DOI: 10.1080/08839514.2022.2031823
- [5] Khaki, S., and L. Wang. 2019. Crop yield prediction using deep neural networks. *Frontiers in Plant Science* 10. doi:10.3389/fpls.2019.00621.
- [6] Cedric, Lontsi Saadio, et al. "Crops yield prediction based on machine learning models: Case of West African countries." *Smart Agricultural Technology* 2 (2022): 100049.
- [7] El-Kenawy, El-Sayed M., et al. "Predicting potato crop yield with machine learning and deep learning for sustainable agriculture." *Potato Research* (2024): 1-34.
- [8] Bi, Luning, and Guiping Hu. "A genetic algorithm-assisted deep learning approach for crop yield prediction." *Soft Computing* 25 (2021): 10617-10628.
- [9] Jabed, Md Abu, and Masrah Azrifah Azmi Murad. "Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability." *Heliyon* (2024).
- [10] Sun, J., L. Di, Z. Sun, Y. Shen, and Z. Lai. 2019. County-level soybean yield prediction using deep CNN-LSTM model. *Sensors (Switzerland)* 19 (20):4363. doi:10.3390/s19204363.
- [11] C. Sagana et al., "Machine Learning-Based Crop Recommendations for Precision Farming to Maximize Crop Yields," 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2023, pp. 1-5, doi: 10.1109/ICCCI56745.2023.10128525.