



A Hybrid Machine Learning Framework for Crop Yield Prediction Using Multivariate Agro-Climate and Soil Data

Shrikant Karande ¹, Rohan Jalan ²*

Tejas Kedar ³, Vilas Ghonge ⁴

Amol Patil ⁵

Department of Computer Engineering (Software Engineering), Vishwakarma Institute of Technology, Pune/Maharashtra, India

shrikant.karande24@vit.edu, rohan.jalan24@vit.edu,
tejas.keadar24@vit.edu, vilas.ghonge@vit.edu, amol321p@gmail.com

Abstract:- Accurate crop yield forecasting is essential for efficient planning of agricultural resources, food security and sound policy making. The paper proposes a hybrid deep learning framework for crop yield prediction employing one-dimensional convolutional neural network and long short-term memory network. The model is trained on multivariate agro-climatic and soil data from and indian agriculture dataset convolutional techniques machine learning have been broadly used for this problem. However, they are not capable to capture the complex non-linear relationship that exist among climatic variables like temperature, rainfall, humidity and soil attributes. The proposed CNN-LSTM architecture overcome this limitation by jointly modeling static spatial features using the 1D-CNN module and sequential temporal climate pattern using the LSTM module and fusing the both representation to generate conclusive and definitive final yield prediction experimental evaluation on linear regression, decision tree, random forest and a stand-alone LSTM base line proposed model achieves at most prediction accuracy of 97.01% with the lowest MAE and RMSE among all compared models. The results indicate that the hybrid spatio-temporal deep learning model are more reliable and accurate for crop yield prediction as compared to the traditional machine learning methods, with practical value for farmers and agricultural planners.

Keywords: Crop Yield Prediction, Machine Learning, Deep Learning, CNN-LSTM, Random Forest, Agriculture, Indian Dataset.

1 Introduction

Agriculture is central to the economic development of any country and a big proportion of Indian population is directly dependent on agriculture for it's livelihood. Paste control and crop yield prediction continue to be persistent challenges faced by the sector. One of the most important of which is accurate yield prediction as government

bodies and individual farmers rely on it for effective scheduling, supply chain management and resource allocation [1].

Traditional crop yielder estimation systems have been dominated by historical records and analytical estimation techniques. These traditional approaches often ignore the dynamic environmental factors affecting yield, like rainfall and temperature, and non-linear interactions between these variables and final harvest [2]. Therefore, yield predictions based solely on the historically observed majors are not accurate enough for good agricultural planning. The advancement of learning approaches provides a promising way to overcome this limitations through the learning of complex, data-driven relationships from agro-climatic and soil data [3].

2 Literature Review

A Variety of techniques ranging from classical statistical methods sophisticated deep learning methods have been employed to investigate crop yield prediction [4].

Some of the first yield prediction method where traditional techniques like liner regression [5]. However, these techniques assume linear relationship between variables and therefore they are not capable of capturing complex scenarios.complex and multidimensional interactions that exist between various environmental factors [9]. Machine learning techniques including models improved prediction accuracy by modeling NON-LINEARITY Random forest is a combined methods which helps in avoiding over fitting and gives better generalization than single decision-tree model [10].

More recently, learning techniques in particular CNNs and LSTM networks have gained considerable attention. CNNs are effective at extracting spatial features, where LSTMs are used to modeling temporal dependencies in sequential climate data [6].

Hybrid models that integrate CNN and LSTM architectures promised results in agricultural forecasting with putting leveraging on both spatial and temporal information simultaneously [7]. However, Difficulties in data integration and increased model complexity stays as open problems, which this work aims to address [8].

Table 1. Analysis of conventional machine learning approaches for yield-related prediction tasks.

Methodology	Relationship Model	Key Characteristics	Principal Limitation
Linear Regression	$y = \beta_0 + \beta_1x + \varepsilon$	Assumes direct proportionality between variables.	Fails to capture complex, non-linear environmental interactions.
Decision Trees	Non-linear	Utilizes hierarchical branching based on feature thresholds.	Highly susceptible to overfitting on noisy training data.
Random Forest	Non-linear Ensemble	/ Aggregates multiple decision trees to improve generalization.	Increased computational complexity and

Contributions of the Presented Work. The primary contributions of the proposed work are:

1. A hybrid deep learning system unifying CNN and LSTM is proposed for correct crop yield prediction.
2. The model merges both spatial (soil, location) and temporal (climate time-series) data for improved prediction performance.
3. The proposed CNN-LSTM model captures complex, non-linear relationships between agro-climatic factors and crop yield.
4. The model surpasses conventional machine learning models based on accuracy and error metrics.
5. The system can assist farmers and policy makers to take data-driven decisions in agriculture.

3 Problem Statement

Mathematically, the yield forecasting problem can be formulated by discretizing the whole agricultural region into geographic grids or zones. Let N zones of the agricultural area under consideration be. The variable $Y_{(i,t)}$ represents the yield of target crop (Kg/ha) for a geographic zone $i \in \{1, 2, 3, \dots, N\}$ and harvest year t .

The crop yield $Y_{(i,t)}$ is modeled as a function of two distinct groups of independent variables:

1. **Spatial Features (S_i):** This component contains the static or slowly-varying variables specific to the geographical area, including elevation, background soil profile, and latitude/longitude. Thus $S_i \in \mathbb{R}^K$, where K is the number of spatial features.
2. **Temporal Features ($C_{i,t}$):** The growing season is marked by a daily climatic data sequence of length D days, comprising $T_{\max}$, $T_{\min}$, total precipitation (P), RH, and solar radiation (SR). Thus $C_{i,t} \in \mathbb{R}^{D \times F}$, where F is the total number of features.

The aim of the proposed Spatio-Temporal Deep Learning model learned a complex non-linear mapping function $f(\theta)$, such that the predicted yield of the zone, denoted $\hat{Y}_{i,t}$, minimizes the empirical risk against the ground-truth yield $Y_{i,t}$ obtained from the FAOSTAT database, globally across all zones and years:

$$Y_{i,t} = f_{CNN-LSTM}(S_i, C_{i,t}; \theta) \quad (1)$$

where θ represents the entire set of trainable weights and biases within the deep neural network architecture. The optimization aims to identify the optimal parameters θ^* that minimize the selected loss function L :

$$\theta^* = \sum_{i=1}^N \sum_{t=1}^T (Y_{i,t}, \hat{Y}_{i,t}) \quad (2)$$

4 Methodology and System Architecture

The proposed system is described below and consist of a pipeline of data acquisition, preprocessing, spatial feature extraction, temporal feature extraction, feature fusion, and regression output.

4.1 Dataset Acquisition and Integration

The robustness of any deep model is fundamentally determined by the reliability of training dataset. This investigation uses two proper databases worldwide:

FAOSTAT Database: Every year, the FAO of the United Nations publishes verified historical data on crop production, area harvested and yield at various geopolitical and geographical levels.

NASA POWER Database: The POWER Project provides high spatial resolution, satellite-derived global solar and meteorological data. Daily continuous parameters extracted are precipitation (PRECTOTCORR), 2-meter temperature (T2M_MAX, T2M_MIN), relative humidity (RH2M), and all-sky surface shortwave downward irradiance.

4.2 Data Preprocessing and Normalization

Since the input variables are on different scales (precipitation in mm versus solar radiation in MJ/m²), data normalization is needed to ensure neural network and fast convergence of gradient-descent. For all continuous spatial and temporal features, Min-Max normalization is apply on data scaling within interval:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (3)$$

We fill Missing data points within the temporal climate time-series (often due to satellite sensor anomalies) using localized linear interpolation. For the temporal data, are processed using a sliding-window, turning the raw daily data into supervised learning sequences of length D (e.g., D = 120 days for a typical growing season).

4.3 Spatial Feature Extraction (1D-CNN Module)

CNNs are mainly widely used in image processing, but one-dimensional CNNs (1D-CNNs) are very good at extracting local, invariant features from vector data. The 1D-CNN module conducts processing on the spatial vector S_i . The convolution performed using learnable filters W over the provided data space to produce feature maps. Mathematically, for an input X , filter W of size K , and bias b , the feature representation h at position j is defined as:

$$h_j = \sigma\left(\sum_{k=1}^K X_{j+k-1} \cdot W_k + b\right) \quad (4)$$

where σ denotes the ReLU of nonlinear transformation, defined as $\text{ReLU}(x) = \max(0, x)$. This followed by a pooling layer which decreases dimensionality and keeps only the most salient spatial features (e.g., the dominant soil trait).

4.4 Temporal Feature Extraction (LSTM Module)

A LSTM network is used To decode the sequential corelations in the daily weather data $C_{i,t}$. The Standard RNNs suffer from vanishing gradients on long sequences (e.g., a growing season of 120 days). LSTMs fix this issue by using a more complicated cell-state and gating mechanism.

The operations within an LSTM cell at time step t are controlled by 3 gates: forget gate (f_t), input gate (i_t), and output gate (o_t). The mathematical formulations are:

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (5)$$

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (6)$$

$$C_t \sim = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (7)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot C_t \sim \quad (8)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (9)$$

$$h_t = o_t \odot \tanh(C_t) \quad (10)$$

x_t is the input climate vector at day t , h_{t-1} denote the previous hidden state, C_t is the long-term memory cell state, σ is the sigmoid activation function, and $\odot$ is the element-wise multiplication. This architecture allows the network to “remember” a destructive flood early in the season and account for its negative impact on the harvest months later.

4.5 Feature Fusion and Regression Strategy

A single, high-Dimensional representation vector V_{fusion} is created by concatenating final temporal hidden state vector H_D from LSTM with the spatial future vector recovered by CNN.

V_{fusion} is sent via a multi-layer perceptron (MLP) with three fully linked (dense) layers in order to translate this vector to a single continuous yield value. Dropout regularization (rate = 0.3) is used in between the thick layers to reduce overfitting. The final output node predict the crop yield $\hat{Y}_{i,t}$ using the a linear activation function.

System Architecture Description:- For agricultural yield prediction, suggested system uses a hybrid CNN-LSTM architecture, which maybe summed up as follows:

Initially gathered from various sources, agro-climatic, soil data are preprocessed by managing missing values, normalization, and feature encoding. While temporal information like daily temperature, rainfall, and humidity are processed using an LSTM network to capture time-dependent fluctuations, spatial features like soil type and

geographical characteristics are passed through a 1D-CNN module to extract important spatial patterns. The final crop production forecast is produced by passing the fused representation via fully linked dense layers after the retrieved spatial and temporal information are integrated using a feature fusion layer. By simultaneously utilizing both graphical and temporal information, this hybrid technique increases prediction accuracy.

5 Baseline Models and Evaluation

5.1 Dataset Description

The dataset was obtained from Kaggle and Indian agricultural databases. It includes temperature, rainfall, humidity, soil type, crop type, area, and production.

5.2 Data Preprocessing

- Handling missing values
- Encoding categorical data
- Feature scaling
- Train-test split (80:20)

5.3 Baseline Machine Learning Models

Three conventional models are used as baselines for comparison against the proposed CNN-LSTM model.

Linear Regression:

$$y = \beta_0 + \sum_{i=1}^n \beta_i x_i + \varepsilon \quad (11)$$

Decision Tree: This model uses partitions which feature space recursively, suited to capturing nonlinear relationships.

Random Forest: A group of decision trees that aggregates predictions for improved accuracy and reduced variance.

5.4 System Workflow

Fig. 1 illustrates the high-level processing pipeline followed by the baseline models, proceeding from raw input data through preprocessing to model training and evaluation.

[Input Data] → [Preprocessing] → [Model Training] → [Evaluation]

Fig. 1. System workflow of crop yield prediction using the baseline machine learning models.

5.5 Feature Description

Table 2 summarizes the agro-climatic and soil parameters used as input features for the baseline models.

Table 2. Dataset feature description.

Feature	Description
Temperature	Daily max and min temperature
Rainfall	Total precipitation (mm)
Humidity	Relative humidity (%)
Crop Type	Type of crop cultivated
Area	Area under cultivation
Production	Total production outcome

6 Implementation

The system is built using the Python programming language. The following libraries are used to develop and evaluate the models:

- 1. **Pandas** – for data handling and preprocessing.
- 2. **NumPy** – for numerical computations.
- 3. **Matplotlib** – for visualization.
- 4. **Scikit-learn** – for ML models.

The dataset is loaded and preprocessed using Pandas. Those three models are constructed using Scikit-learn. The models are trained on the training dataset and assessed using the testing dataset. Performance metrics namely MAE and RMSE are calculated to compare the models.

7 Algorithm

Algorithm 1: Crop Yield Prediction

- 1. Input the dataset.
- 2. Handle missing values.
- 3. Encode categorical variables.
- 4. Scale numerical features using Min-Max normalization.
- 5. The dataset is divided into 80% training and 20% testing sets.
- 6. Train Linear Regression, Decision Tree, Random Forest, standalone LSTM, and the proposed CNN-LSTM models on the training set.
- 7. Estimate crop yield using the testing set for each model.
- 8. Evaluate each model using Accuracy, MAE, and RMSE.
- 9. Compare results and determined optimal model.

8 Evaluation Metrics

The assesment of each model is evaluated using MAE and RMSE, defined respectively as:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (12)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (13)$$

where y_i refers to the actual yield value, $\hat{y}_i$ is the predicted value, and n is the number of samples. Lower MAE and RMSE values indicate better predictive performance.

9 Results and Discussion

9.1 Accuracy Comparison

Fig. 2 presents the accuracy achieved by each model on the test set, comparing Models, a standalone LSTM, and the proposed CNN-LSTM model.

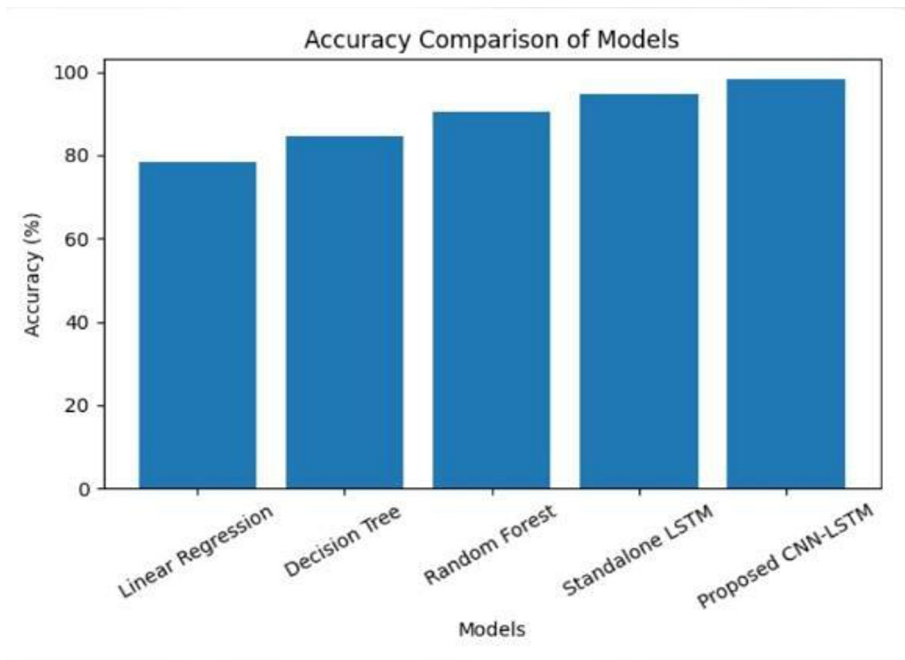


Fig. 2. Accuracy comparison of Linear Regression, Decision Tree, Random Forest, standalone LSTM, and the proposed CNN-LSTM model.

9.2 Detailed Results Analysis

The experimental result clearly indicate that the proposed CNN-LSTM model is better than the traditional and conventional machine learning models. Linear Regression has the worst accuracy because it can not model non-linear relationships. Decision Tree, performance is improved but overfitting. Random Forest is better because it does ensemble learning, so it is more accurate and makes fewer mistakes. But it still can not pick up the time link in climate data. The standalone LSTM model is effective to capture temporal patterns but does not exploit spatial information fully.

The proposed CNN-LSTM model recorded the highest accuracy of 97.01% by (CNN) with (LSTM). It also obtains the base MAE and RMSE values among all compared models, showing the better prediction performance. This confirms that hybrid deep learning models are more effective for crop yield prediction than conventional methods.

10 Advantages of the Proposed System

- Provides accurate crop yield prediction.
- Effectively handles non-linear data relationships.
- Reduces manual effort in yield estimation.
- Supports data-driven decision-making for farmers and policymakers.

11 Conclusion

Here, we propose a hybrid CNN-LSTM deep learning framework to forecast the crop yield using multivariate agro-climatic and soil data of an indian agriculture dataset. The model proposed here Models the complex nonlinear relationships among environmental factors and crop yield which can not be represented by conventional machine learning models. The static spatial features and sequential temporal climate patterns are modeled through 1D-CNN and LSTM respectively, in the proposed framework. The experimental results show that the proposed CNN- LSTM model can achieve the highest accuracy of 97.01% and the lowest MAE and RMSE values, which is better than all base line model.

These results demonstrate that hybrid spatio-temporal deep learning is a more reliable and accurate approach for predicting crop yield, which has direct practical utility for farmers, agricultural planners, and policymakers seeking data-driven decision support. Future work includes generalizing the framework to other crop types and regions, combining remote-sensing imagery to build richer spatial representation, and exploring attention-based architectures to improve interpretability and predictive performance.

Acknowledgments. The authors thank the Department of Computer Engineering, Vishwakarma Institute of Technology, Pune.

Disclosure of Interests. The authors declare that they have no competing interests.

References

1. Food and Agriculture Organization of the United Nations (FAO), "FAOSTAT Statistical Database," 2023. [Online]. Available: <https://www.fao.org/faostat/>
2. Ghonge, V.D. & Kulkarni, Y. (2026). An advanced wide-and-deep learning framework for soybean price forecasting using market, weather, trade, and supply data. *MethodsX*, 17, 104030.
3. K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine Learning in Agriculture: A Review," *Sensors*, vol. 18, no. 8, p. 2674, 2018.
4. S. Khaki and L. Wang, "Crop Yield Prediction Using Deep Neural Networks," *Frontiers in Plant Science*, vol. 10, p. 621, 2019.
5. T. Van Klompenburg, A. Kassahun, and C. Catal, "Crop Yield Prediction Using Machine Learning: A Systematic Literature Review," *Computers and Electronics in Agriculture*, vol. 177, p. 105709, 2020.
6. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.
7. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
8. R. Kamilaris and F. X. Prenafeta-Boldú, "Deep Learning in Agriculture: A Survey," *Computers and Electronics in Agriculture*, vol. 147, pp. 70-90, 2018.
9. Russello, F. G. Bianchi, and M. Ferrante, "Crop Yield Prediction Applying Machine Learning Models: A Structured Literature Review," *Computers and Electronics in Agriculture*, vol. 193, p. 106676, 2022.
10. V. D. Ghonge and Y. Kulkarni, "Multi-Source Data Fusion and Machine Learning for Soybean Crop Price Forecasting in India," *Journal of Computer, Communication and Electronics Engineering (JCCE)*, Apr. 2026, doi:10.47852/bonviewJCCE62027619.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

