RESEARCH PAPER

Evaluating Machine Learning Models for Pan-Indian Agricultural Yield Prediction

[1]Tambe S.L.*and [2]Dr.Dashore Pankaj

[1]Research Scholar School of Computer Sciences and Engineering, Sandip University Nashik (MH), India.
[2]Professor School of Computer Sciences and Engineering, Sandip University Nashik (MH), India
Email id: [1]tambesl.comp@gmail.com, [2]dashorepankaj@gmail.com

Received: 16th Dec, 2025; Revised: 8th Feb 2026; Accepted: 12th Feb, 2026; Available Online: 28th Feb, 2026

ABSTRACT
This paper presents a comparative evaluation of the classical machine learning models and the deep machine learning models in the prediction of the crop yield using a pan-Indian agricultural dataset. A government-based composite dataset comprising 19,690 samples was built and a strict feature engineering pipeline was adopted to deal with quality of data and multicollinearity. There were four models namely Linear Regression (LR), Decision Tree (DT), Artificial Neural Network (ANN), and a Convolutional Neural Network – Long Short Term Memory (CNN-LSTM) hybrid that were systematically tested. However, findings showed that the Linear Regression simple model gave a better predictive power with a score of R^2 of 0.9942. It concludes the paper by finding that in the case of structured tabular data at national scale computationally efficient classical models combined with careful feature engineering can be more effective at crop yield forecasting than more complicated deep learning methods.
Keywords: *Machine Learning (ML), Linear regression (LR), Decision Tree (DT), Artificial Neural Network (ANN), Convolutional Neural Network – Long Short Term Memory (CNN-LSTM)*
How to cite this article: Tambe SL and Dashore P, Evaluating Machine Learning Models for Pan-Indian Agricultural Yield Prediction. Int J Drug Deliv Technol. 2026;16(3): 747-756. DOI: 10.25258/ijddt.16.3.82
Source of support: Nil.
Conflict of interest: None

INTRODUCTION

The agricultural sector has been the foundation of the economic structure in India and it has remained a key determinant of subsistence and national well-being. This is a very crucial sector that has earned millions of people economic stability and at the same time ensured that there is nutritional stability in the country over the generations [2]. In addition to making its direct contributions, the agriculture sector supports a large chain of secondary industries and is the economic base of rural populations throughout the country.

The agriculture systems in India are progressively becoming vulnerable to the ever-increasing effects of the climate change that are radically changing the traditional ways of farming in the country. The changing climatic conditions with erratic precipitation patterns and increased occurrence of extreme weather outbreaks is disrupting agricultural production by causing change in the growing seasons and reduced crop yields. This volatility of the environment targets marginal cultivators who have low access to capital and technology and are therefore limited in their ability to adapt. Moreover, such climate-related stresses are combined with enduring agricultural limits, such as declining soil health, shrinking water supplies and changing biotic stressors with all of them challenging the sustainability of the farming ecosystems in the long term [5].

The rapidly increasing method is machine learning which assists in predicting crop yield. The key concept of deploying machine learning models is that the agriculture industry should be capable of enhancing its output[4]. This would focus on precision agriculture whereby the desirable environmental factors would be balanced with the quality assurance. Crop yield forecasting assists the farmer to determine advance what to grow and at what time to grow.

Crop yield forecasting is a very important task that we undertake in this research in which we use regression algorithms which includes RF, LR, Artificial Neural Networks (ANNs) and CNN-LSTM which are well known to be effective in prediction modeling. RF is ML algorithm that is an ensemble based on the concept of Bootstrap Aggregation (Bagging) to enhance predictive accuracy. In the process of training, a number of decision trees are created as base learning models and the final prediction is an average of the projections made by every decision tree. The independent parallel characteristic of decision trees in RF leads to the strong predictions with lesser variance. ANNs are models of computing that are based on biological concepts neural networks. ANNs are made up of networked nodes (neurons) in a layer: input, hidden and output layer. Their ability to represent nonlinear relationships makes them especially useful in solving complex problems including crop yield prediction which is

**Author for Correspondence: tambesl.comp@gmail.com*

able to process high-dimensional data and do it in parallel. We are mainly focused on determining the performance of these algorithms with respect to forecasting crop yields with reference to pre-processed datasets. In particular, we pay attention to the analysis of the Mean Squared Error. The predictive accuracy of the three techniques was to be compared using (MSE), R-squared (R^2) and Mean Absolute Error (MAE) measures.

Main input is the comparative analysis and confirmation of four different machine learning frameworks LR, DT Regression, ANN and hybrid CNN LSTM framework specifically designed to predict crop yields based on structured agricultural dataset. The analysis of the performance of each model systematically provides the Linear Regression model as the best option to use in this task with the highest R^2 score of 0.9942 and the lowest error values (MSE 4629.04 and MAE 7.88). This finding shows that a model with lower complexity which is computationally efficient can be more effective than more complicated models such as the CNN LSTM (R^2 0.8669) on the presented data and successfully strikes the right balance between the high predictive performance and working speed.

The suggested system combines information on Indian OGD, FAO, and IMD, which contains 19,690 samples of various crops, seasons and agroclimatic regions. One of its major innovations is that it has a feature engineering pipeline in which it uses logarithmic scaling to create interaction terms such as Productivity and Fertilizer Efficiency to understand the complex agronomic relationships. Comparison of four models such as ANN and CNN-LSTM architectures is made. The presented hybrid method allows to model both the spatial and temporal trends and offers a sound instrument of optimizing the resources and agricultural planning in all the diverse farming landscapes of India.

LITERATURE SURVEY

Another study by Wang Y et al. [1] is focused on the advances in crop yield prediction in cases when deep learning algorithms are used and their advantages are contrasted with are traditional machine learning algorithms that tend to produce less accurate results. It discusses several models and approaches of prediction and examines their prospects in deep learning to achieve sound predictions in agriculture. This supports the importance of crop predictions and the effects of crop predictions on making a decision by a farmer and economic policy. Another beneficial aspect of the study is future direction of pure deep learning-based crop yield prediction systems through the enumeration of limitations of the existing systems and the next path they should take. Nigam A. et al. [2] Machine learning has been intrudingly adopted in precision agriculture which enables one to have more accurate crop yield prediction with data of historical records, real time sensing data and satellite imagery. Advanced algorithms, such as regression models and deep learning techniques, offer the ability to model the complex environmental and agronomic interactions resulting in the ability to estimate the yield in a way that is accurate. Combined sensor and satellite data can be used to make better decisions in order to control resources e.g. water and fertilizers and ensure that people adopt sustainable agricultural methods. Even though data quality and model interpretability are still being considered the center of attention in future Research, the developments serve the essential purpose of tackling the issues of climate variability and food safety. One such study was by Savaliya L et al. [3] their ensemble model, which included K-Means clustering, RF, LR and SVM was used in prediction of crops with an accuracy of approximately 99.77%. The model relies on majority voting strategy to make the prediction reliable by leveraging the strengths of various algorithms. The blended strategy integrates the environmental state together with the site conditional information to prescribe cropping according to the information. This study notes that ensemble learning is significant in enhancing crop productivity as well as augmenting economic returns of farming activities. The introduction of ensemble techniques therefore has turned out to be a transformative and strong force that will help in advancing precision agriculture.

To test the gap in rice yield in Eastern India under climatic variability Sahoo, Singha and Govind (2024) tested several machine learning algorithms such as Cubist, RF, gradient boosting machine (GBM), multivariate adaptive regression splines (MARS), SVM and extreme gradient machine learning (XGB) [4]. Their research used climatic and soil variables and results of the models were cross validated with 1,889 field observations. The tested models showed that Cubist and RF models had better predictive accuracy and the rest of the algorithms showed low performance. Jabad and Murad (2024) have conducted a systematic literature review on the subject of applying the ML and DL methods to crop yield forecasting. Their analysis compared the work of the algorithms including RF, ANN, SVM, CNN-LSTM networks and DNN. The authors have determined the most important environmental predictors such as precipitation, soil characteristics and thermal conditions and vegetation indices (NDVI, EVI, LAI) as the key inputs in model training [5]. Noma and Babu (2024) introduced a machine learning model which is built on Optimized Gradient Boosting to predict the use of Climate Smart Agriculture (CSA) practices. The researchers used longitudinal household survey data in the Rakai district of Uganda and included in the analysis socio-economic indicators, agroecological conditions and climatic variables as predictive variables. Their model had a moderate impact of prediction accuracy of 60 which shows that it has potential in identifying trends of CSA adoption [6].

The study by Talaat F. (2023) introduces a new Crop Yield Prediction Algorithm (CYPA) framework involving the delivery of machine learning approaches with Internet of Things (IoT) technology. The paper comparatively analyzes five regression models DT, RF, and Extra Trees regressors that are optimized by virtue of hyper parameter

tuning. Excellent predictive performance was proved by the experimental results and the model accuracy scores were 0.9814, 0.9933, and 0.9903 respectively. Moreover, the authors have added the active aspect of learning to the CYPA system that greatly increased the computational and predictive efficiency of the system in agricultural yield forecasting scenarios [7]. In the study by Srivastava et al. (2022), a 1D-CNN model was offered to make predictions of winter wheat yields in 271 German counties during an interval of 1999-2019. The model took into consideration the weather, soil data and phonological data. It performed better than eight baseline models such as RF, XGBoost, and DNN by lowering RMSE by 7-14% as well as improving correlation by 4-50%. SHAP analysis showed that the non-random variables that contributed to significant predictions included DUL, wind speed in week 10 and radiation in week 7. Though the model was very accurate constraints were observed in the model namely black box characteristics and errors whose regional variations were mentioned to be [8]. Van Klompenburg et al. (2020) performed a Systematic Literature Review (SLR) study to investigate the use of machine learning in crop yield prediction. They accessed 567 studies in six databases and 50 of them were selected according to inclusion/ exclusion criteria. The important features (temperature, rainfall, soil type) and algorithms (e.g., Artificial Neural Networks) were obtained. An additional search found 30 studies based on deep learning, which analyzed CNN, LSTM, and DNN [9]. This study was conducted by X. E. Pantazi et al. [10] Using machine learning methods and modern remote sensing technologies to predict wheat production. According to results we find that predictive algorithm implementation along with spatial sensing information can be effective avenues to enhancing precision in the estimations of yields and the allocation of resources in wheat culture. Holzman et al. [11] Study considers a pre-crop analysis of crop productivity grounded on remotely sensed water stress and sun radiation. This implies that remote sensing can be employed in forecasting the performance of crops and alleviating risks involved in the management of agriculture. In this study Singh et al. [12] discuss application of machine learning in High throughput stress phenotyping of plants. It also investigates stress variables of crops to give information about enhancing agricultural resilience and yield using advanced phenotyping technologies. The Research employs non-linear parametric modeling in order to estimate the impacts of the soil properties on crop yields and NDVI among others [13].

The SMART-CYPS model created by Kuradusenge and colleagues (2024) is an IoT device that integrates machine learning to forecast crop harvests in sub-Saharan Africa. The system predicted potato and maize with an MAPE ranging between 0.177-0.339 by measuring the weather and soil conditions in Rwanda, Musanze District. Although it produces positive forecasting results, the authors suggest an increase in the system in terms of crop coverage and geographical applicability in order to contribute to the food security further [14]. Khanagoudar et al. (2023) conducted an extensive systematic review of the machine learning use in crop yield prediction through climatic factors. Their comparison considered various regression algorithms such as decision trees, k-nearest neighbors, random forests and linear regression models. The article used critical predictive variables like meteorological factors and plant pathological variables [15]. article by Bloh, Lobell, and Asseng was a machine learning system of wheat yield that integrate DSSAT simulations with neural networks and random forests into knowledge. Their hybrid method produced synthetic training data using crop and climate models, 8% smaller in prediction error than traditional methods and proving the usefulness of process-based model with machine learning to create climate-resilient agriculture [16].

There are significant limitations to existing systems. To begin with, numerous researches are based on limited datasets that are characterized by low crop diversity or geographic range, minimizing the level of generalization. Second ensemble models and deep learning models are highly accurate but tend to have black-box properties which are not interpretable, which prevents trust in the model by the stakeholders. Third the critical data quality problems such as multicollinearity of the input features and local prediction errors are poorly addressed. Fourth majority of the models do not extensively feature engineer to reflect the intricate agronomic interaction. Lastly, the current methods have a small combination of different data sources and lack of validation in different agro-climatic regions and limiting their applicability to large-scale agroforestry decision-making.

METHODS AND MATERIALS

This work is based on a composite dataset which were found in Indian Open Government Data (OGD) Platform, UN Food and Agriculture Organization (FAO) and Indian Meteorological Department. The work on Google Colab was done to merge data and pre-process it. The dataset used for this research contains 19690 sample and 10 features. The features gives us following information:

Crop : Name of the crop cultivated, Crop_year : Year of crop cultivation, Season : agriculture season (kharif/rabi), state : Indian states where crop was cultivated, Area : Cultivated area for the crop, production : Total production of the crop in tons, Annual rainfall : Total annual rainfall in the region, Fertilizer : Total fertilizer used, Pesticide : Total pesticide used and Yield : Crop yield per unit area. The key feature of this dataset is it covers multiple crop that includes cereals, pulses, oilseeds, fibers, spices, and horticultural crops. This dataset contains regional diversity among the states from different agro climatic zones. Its contains input and output metric of area, production, yield and input factors such as fertilizers and pesticides.

EXPLORATORY DATA ANALYSIS (EDA):

Exploratory Data Analysis (EDA) is the diagnostic step as the data-based research prerequisite of any data-driven research especially regarding complicated agricultural data. Prior to the implementation of machine learning

algorithms in predicting the yield, EDA gives a meaningful pictorial information about the data structure, quality, trends and relationships that have a direct impact on the selection of model, feature engineering and results interpretation.

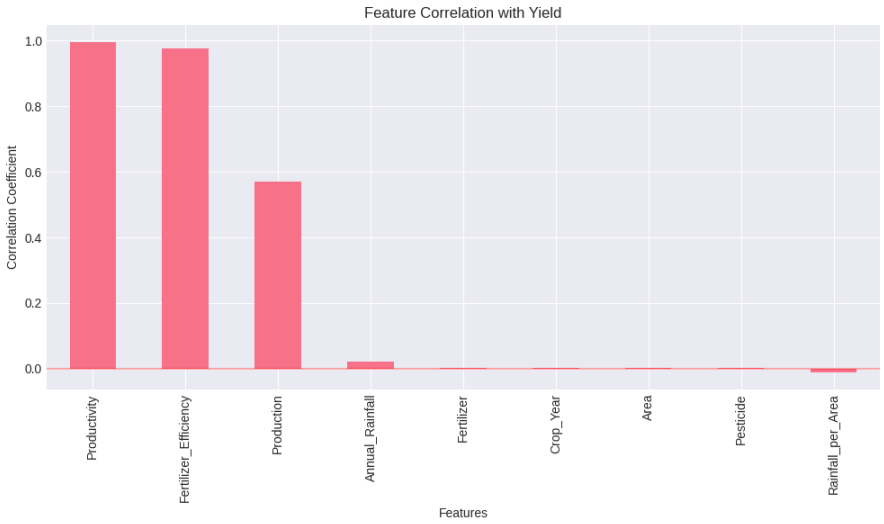

Figure 01: Feature Correlation with yield

Figure 01 shows the results obtained after performing Exploratory Data Analysis (EDA) on the dataset. From the above figure 1 we can conclude that yield is strongly correlated with productivity, Fertilizer Efficiency and production features. This strongly suggest that yield prediction should prioritize these features.

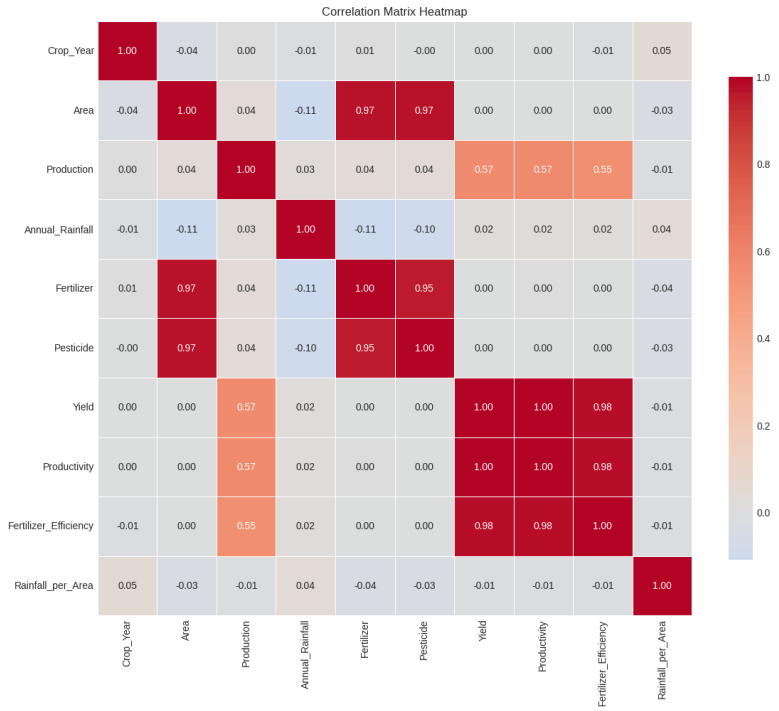

Figure 02: Correlation matrix heatmap

Figure 2 demonstrates an almost perfect relationship (r= 0.95 to 0.97) between Area and Fertilizer and Pesticide is a sign of severe multicollinearity which shows the actual connection of the world in which the bigger the farm is the higher the use of inputs. This interdependence in regression distorts coefficient estimates inflates the standard errors and increases the chances of overfitting resulting in less reliable and generalizable models. Application of feature engineering on any model has more opportunities to do better. The two factors have an influence on performance of any model. Pre-processing and manipulation of the data. Feature engineering includes manipulation of the information to enhance the precision of the model.

After a thorough data preprocessing phase, a powerful feature-engineering pipeline was implemented in order to enhance the predictive power of the dataset as far as agricultural yield modeling is concerned. A total of nine new features were systematically created through the pipeline and this increased the data matrix to 22 columns. The main changes included the use of logarithmic scaling (log_Area, log_Production, log_Fertilizer, log_Pesticide) to rectify skewness of salient agronomic variables with the resultant stabilization of variance and normalization of distributions thereof. Two terms of interaction were added to help gain synergies namely Rainfall_Area_Interaction and Fert_Pest_Interaction. Also a nonlinear, quadratic, polymorphic attribute, Rainfall_squared was created to cipher the possible nonlinear connection between annual rainfall and crop yield.

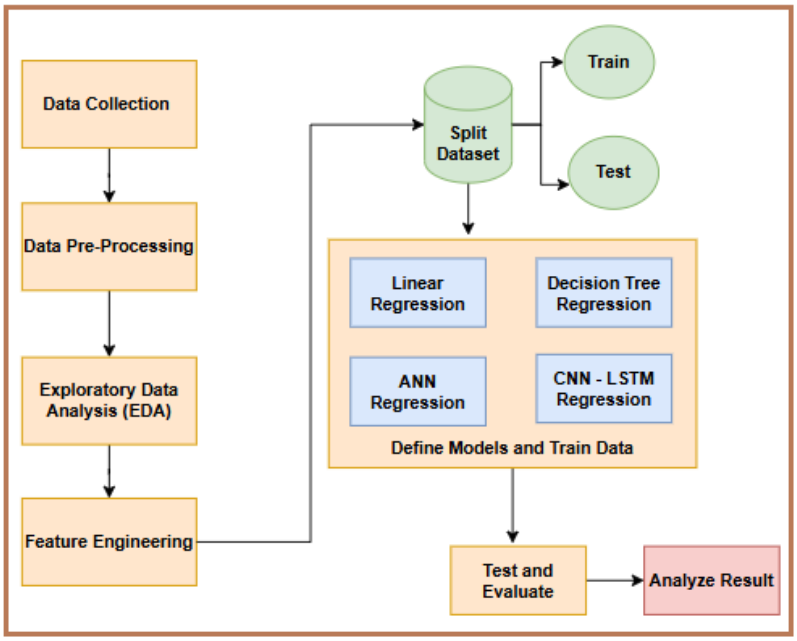

Figure 03: Proposed Methodology

Figure 3 shows the proposed methodology of the research. The first step involves gathering the data in the different sources as explained in above. The perform EDA in order to obtain the insights of the dataset to train machine learning model appropriately. Then we divided the dataset into test and train. This model is constructed with the help of LR, DT regression, ANN and CNN-LSTM regression. The last step is to measure the outcome of the model.

Flow of the algorithms

Let $D = \{(X_i, Y_i)\}^N{}_{I=1}$ be the raw dataset, where each $X_i \in R^d$ represents agricultural features and $y_i \in R$ is the corresponding crop yield.

Input: Raw dataset $\mathcal{D}$, target variable name 'Yield', Output: Performance metrics for all models, predictions.

Step 1: Initialize Preprocessing

For each feature $X^j \in X_i$:

if $x^{(j)}$ is numerical:

if $x^{(j)}$ is null:

$x^{(j)} \leftarrow \text{median}\{x_i^{(j)}\}^N{}_{i=1}$

else if $x^{(j)}$ is categorical:

if $x^{(j)}$ is null:

$x^{(j)} \leftarrow \text{mode}\{x_i^{(j)}\}^N{}_{i=1}$

Step 2: Feature Engineering for i=1 to N

$p_i = y_i / (\text{Area}_i + \varepsilon)$ // Productivity

$f_i = y_i / (\text{Fertilizer}_i + \varepsilon)$ // Fertilizer Efficiency

$r_i = \text{Rainfall}_i / (\text{Area}_i + \varepsilon)$ // Rainfall per Area

Step 3: Model training Sequence

Split D into D_{train} (80%) and D_{test} (20%)

For each model m $\in$ {LR, DTR, ANN R, CNN-LSTM R}

Initialize model parameters θ_m

For t = 1 to T_{max} (or until convergence):

If m == LR:

$\theta\{LR\} \leftarrow \text{argmin}_\theta \sum\_\{(x,y)\in\mathcal{D}\_train\} (y - (w^{Tx+b}))^2$

Else if m == DT:

Recursively partition feature space to minimize:

$\sum\_\{node\} \sum \{I \in node\} (y_i - \mu_{node})^2$

Else if m == ANN:

For layer l = 1 to L-1:

$h^{(l)} = ReLU(W^{(l)}h^{(l-1)} + b^{(l)})$

Apply Dropout ($h^{(l)}$, p=0.3)

Apply Batch Norm ($h^{(l)}$)

$y\_pred = W_L h^{(L-1)} + b^{(L)}$

Update θ via back propagation to minimize MSE

Else if m == CNN-LSTM:

// create sequences of length s

For t = s to N_{train}:

$X_{seq} = [x_{t-s+1}, ..., x_t]$

// CNN Feature Extraction

For conv_layer = 1 to 2:

C = Conv1D (X_{seq}, filters=64, kernel=3)

C = MaxPool1D(C, pool_size=2)

// LSTM Temporal Processing

For lstm_layer = 1 to 2:

$h_t, c_t = LSTM(C, h_{t-1}, c_{t-1})$

$y\_pred = Dense(h_t)$

Update θ via back propagation

Step 4: Evaluation Loop

Initialize results dictionary R= { }

For each trained model m:

Calculate:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

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

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

R [m] = {MSE: MSE_m, RMSE: $RMSE_m$, MAE: MAE_m, R^2: R^2_m}

For each model R[m]

Plot performance comparison

The Artificial Neural Network (ANN) uses a four-layer feedforward architecture where it has 481 total nodes with ReLU activations in it. The CNN-LSTM hybrid model by comparison has two layers of convoluting features to extract spatial features, two layers of LSTM to model time variations and two layers of dense which amounts to about 1,250 parameters. The two models received an initial learning rate of 0.001, a batch size of 32 and the same dropout rate (0.3, 0.3, 0.2) and batch normalization. The ANN was programmed to run up to 100 epochs, whereas the CNN-LSTM was configured to process up to 50 epochs as it is more complex in terms of the sequential processing.

RESULT AND DISCUSSION

A variety of standard performance measures have been used to assess the effectiveness of a predictive model such as Mean Square Error (MSE), Mean Absolute Error (MAE) and R-squared (R^2). These evaluation criteria are critical in the context of deep learning applications since they help ascertain the model predictive accuracy.

Mean Square error (MSE), which is explained in the equation below is a measure of the closeness of a regression line to a set of data points. Mathematically it is a risk measure which is the expected squared error loss.

$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ Where y_i = observed Value and $\hat{y}_i$ = predictive value

R-squared (R^2) is indicates how much of the variance is explained by the independent variable as given bellow.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

In this research we implemented LR, DT Regression, ANN Regression and CNN-LSTM Regression model for crop yield prediction.

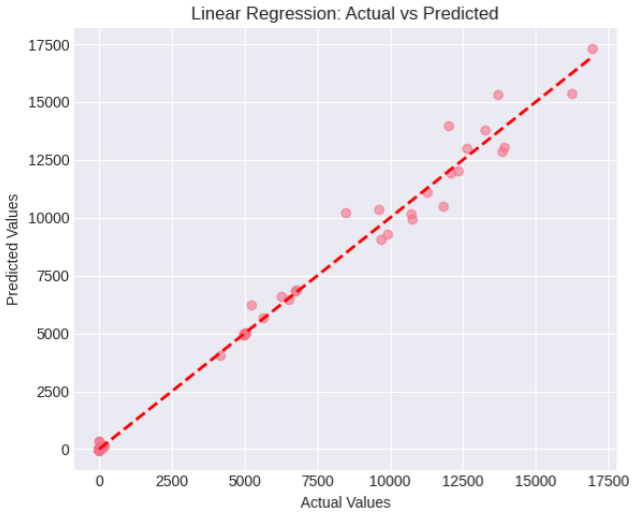

Figure 04: scatter plot for linear regression

According to the visualization and performance measures as shown in above figure 4 it is clear that the linear regression model is almost perfectly fitted. The co-linearity of the data points on the best prediction line on the plot is statistically validated by the fact that the R^2 score of 0.9942 is extremely high that is the model accounts to 99.42 percent of the data variance. The minimum MAE of 7.88 indicates very precise predictions of the points. Although squared error sensitivity increases the value of the MSE and RMSE (4629.04 and 68.04 respectively) both values are lower than the scale of data which in aggregate confirms the strong prediction ability of the model.

Figure 05: scatter plot for Decision Tree regression

Decision Tree Regressor is a model that has a high level of predictive control as shown in above figure 5. The model explains more than 99% of the variation of the target variable which is shown by an R^2 score of 0.9919. The very low MAE of 6.32 indicates very low average error of the real and predicted values and indicates a high degree of precision. Although the MSE of 6501.69 and its derivative RMSE of 80.63 are relatively elevated which indicates the sensitivity of the model to few bigger errors the aggregate statistics affirm the great and precise fitting of the data.

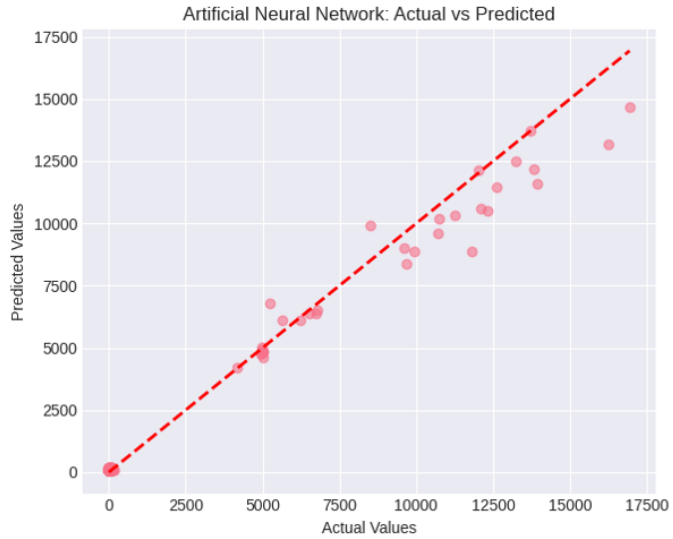

Figure 06: scatter plot for ANN

The Artificial Neural Network (ANN) model as shown in above figure 06 has a good fit and this is evidenced by the good R^2 score of 0.9778 which accounts to the variance of about 97.8% of the data. The metrics of the error demonstrate a more subtle performance. The average prediction deviation of 78.39 is moderately low, with a value of the Mean Absolute Error (MAE). The much greater Mean Squared Error (MSE) value of 17751.15 is evidence that there are instances of higher but more infrequent prediction errors which are more severely penalized by these squared measures. The model has high explanatory potential and a disposition towards periodic errors.

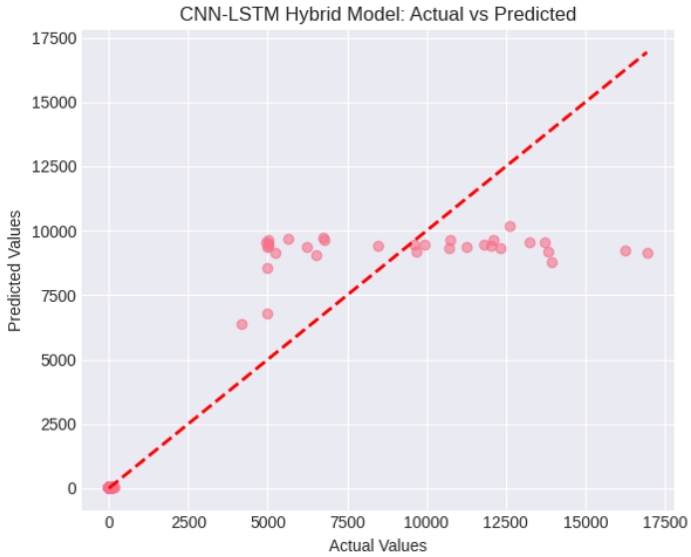

Figure 07: scatter plot for CNN-LSTM

The CNN-LSTM hybrid model as shown in above figure 7 has exhibited a good fit as reflected in the high R^2 score of 0.8669 which appears to cover more than 87% of the data variation. The average error of prediction is reasonable as the MAE is 50.49. The high values of MSE and RMSE mean existence of significant outlier errors that indicate that there is some inconsistency in the performance of the model.

Table 1: Performance Comparison table

Model Name	MSE	RMSE	MAE	R^2
Linear Regression	4629.0388	68.0370	7.8849	0.9942
Decision Tree	6501.6918	80.6331	6.3197	0.9919
ANN	17751.1475	133.2334	78.3928	0.9778
CNN-LSTM	106873.3641	326.9149	50.4909	0.8669

According to the comparative analysis as shown in above table 1 LR showed the most excellent overall performance with excellent R^2 score of 0.9942 and minimum RMSE of 68.04. The lowest MAE of 6.32 was got with the Decision Tree Regressor which provides better average point accuracy. The deep learning models had more error rates. Whereas the ANN was very strong in R^2 of 0.9778, it was much higher in MAE of 78.39. It is interesting to note that CNN LSTM hybrid model showed the worst metrics on all metrics as it had the most errors and lowest R^2 of 0.8669 which indicated that it was not appropriate in this particular case of regression.

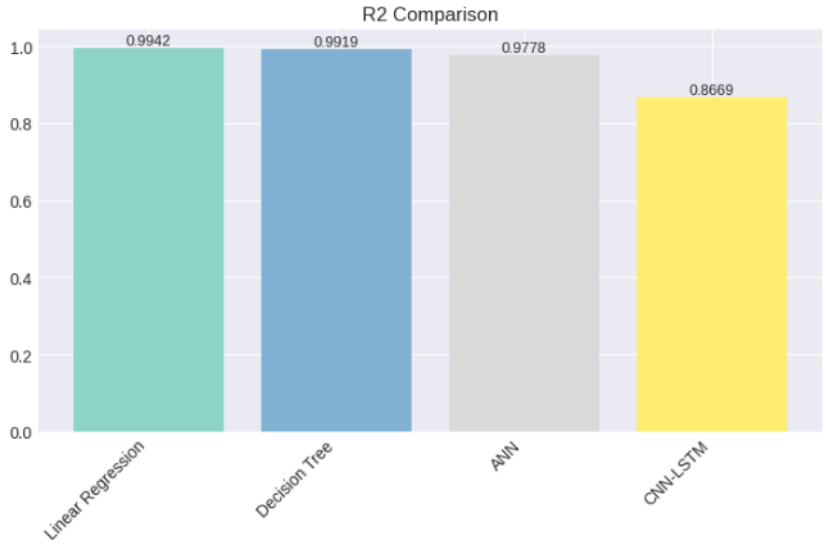

Figure 08: Comparative analysis using R^2

The comparative analysis shows that there is a distinct sequence of model performance as far as crop yield prediction using the composite Indian dataset are concerned. The more elaborate deep learning models were greatly surpassed by classical machine learning models namely Linear Regression and Decision Trees. This finding is consistent with those of Wang Y et al. [1], who described that deep learning does not necessarily ensure high performance in any agricultural setting and its performance is highly data-sensitive.

The strength of the proposed methodology is that it is based on the feature engineering as the cornerstone, which directly controlled the limitations of data quality and multicollinearity as reported in literature [2, 15]. The model focused on the features that were found to be highly empirically correlated with yield and thus it gave up on the simplistic use of environmental and input variables as was common to other models [4, 5]. This preprocessing step probably gave the less complex easier-to-understand models (LR and DT) the ability to produce the underlying relationships at an efficient rate which achieved the highest levels of R^2 scores (greater than 0.99) and low-error scores. ANN and CNN-LSTM models although able to model complex non-linear interactions like investigated by Van Klompenburg et al. [9] and Jabad and Murad [5], were unable to cope with such a tabular non-sequential data. They have a greater MSE and RMSE which means that they are sensitive to outliers and overfitting which has been reported during the study of ensemble and hybrid models [3, 8].

This study shows that in the case of the particular task of predicting yields based on curated national-scale tabular data, simpler, well-tuned algorithms that run on a strictly preprocessed dataset can be even more accurate and reliable than more computationally-intensive deep learning models that providing a feasible and effective alternative to more sophisticated agricultural decision support systems.

CONCLUSION

The results of the research show that simpler machine learning models can be used to make high predictive accuracy on crop yield when tabular data is well prepared. Linear Regression (R^2 = 0.9942) and Decision Trees (MAE = 6.32) produce superior results which proves that basic algorithms perform quite well when paired with extensive preprocessing and human-based feature development. More complex architectures like ANN and Convolutional Neural Network Long Short-Term Memory (CNN LSTM) hybrids were found to be more susceptible to overfitting and generated higher error values. As a result it is more appropriate to devise solid forecasting technologies in agriculture by addressing data refinement and smart design of features than by engaging more complex models.

REFERENCES

1. Wang Y., Zhang Q., Yu, F. Zhang, N., Zhang, X., Li, Y. & Zhang, J. (2024). Progress in Research on Deep Learning-Based Crop Yield Prediction. Agronomy, 14(10), 2264.

2. Nigam A., Garg S., Agrawal A., & Agrawal P. (2019, November) Crop yield prediction using machine learning algorithms. In 2019 Fifth International Conference on Image Information Processing (ICIIP) (pp. 125-130). IEEE.

3. Savaliya L., Sapovadiya M., Garg D., Patel P., & Shah M. (2024, October) Agricultural Analysis of Machine Learning Algorithms for Crop Prediction. In 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) (pp. 1196-1203). IEEE.

4. Sahoo S., Singha C., & Govind A. (2024). Advanced prediction of rice yield gaps under climate uncertainty using machine learning techniques in Eastern India. *Journal of Agriculture and Food Research, 18*, 101424

5. Jabeed M. A. and Murad M. A. A. (2024). Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. *Heliyon, 10*, e04836.

6. Noma F. and Babu S. (2024). Predicting climate smart agriculture (CSA) practices using machine learning: A prime exploratory survey. *Climate Services, 34*, 100484

7. Talaat F. M. (2023). Crop yield prediction algorithm (CYPA) in precision agriculture based on IoT techniques and climate changes. *Neural Computing and Applications*, 35, 17281–17292

8. Srivastava A. K., Safaei N., Khaki S., Lopez G., Zeng W., Ewert F., Gaiser T., & Rahimi, J. (2022). Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Scientific Reports, 12*(1), 32315.

9. van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture, 177*, 105709.

10. X. E. Pantazi, D. Moshou, T. Alexandridis, R. L. Whetton, and A. M. Mouazen, ‘‘Wheat yield prediction using machine learning and advanced sensing techniques,’’ Comput. Electron. Agricult., vol. 121, pp. 57–65, Feb. 2016.

11. M. E. Holzman, F. Carmona, R. Rivas, and R. Niclòs, ‘‘Early assessment of crop yield from remotely sensed water stress and solar radiation data,’’ ISPRS J. Photogramm. Remote Sens., vol. 145, pp. 297–308, Nov. 2018.

12. A. Singh, B. Ganapathysubramanian, A. K. Singh, and S. Sarkar, ‘‘Machine learning for high-throughput stress phenotyping in plants,’’ Trends Plant Sci., vol. 21, no. 2, pp. 110–124, Feb. 2016.

13. R. Whetton, Y. Zhao, S. Shaddad, and A. M. Mouazen, "Non-linear parametric modelling to study how soil properties affect crop yields and NDVI," Comput. Electron. Agricult., vol. 138, pp. 127–136, Jun. 2017

14. Kuradusenge, M., Hitimana, E., Mtonga, K., Gatera, A., Habiyaremye, J., Ngabonziza, J., Hanyurwimfura, D., Rukundo, P., & Mukasine, A. (2024). SMART-CYPS: An intelligent internet of things and machine learning powered crop yield prediction system for food security. *Discover Internet of Things*.

15. Khanagoudar, P. S., Sushma, B. S., Chandana, C. N., Manikanta, N., & Awatimath, S. S. (2023). Crop yield prediction using machine learning algorithm based on climate variables. *International Journal of Cheminformatics, 1*(2).

16. Von Bloh M., Lobell, D., & Asseng, S. (2024). Knowledge informed hybrid machine learning in agricultural yield prediction. *Computers and Electronics in Agriculture, 227*, 109606.