CROP YIELD ANALYSIS AND FORECASTING

Navneet Kumar Singh, B.Tech Information Technology, 4th Year, Government College of Engineering and Ceramic Technology, Kolkata

Saksham Asopa, B.Tech Information Technology, 4th Year, Government College of Engineering and Ceramic Technology, Kolkata

Project Guide Name: Ankit Lodh

Period of Internship: 19th May 2025 - 15th July 2025

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

1. Abstract

This project aims to analyze and forecast crop yields by leveraging historical agricultural data and advanced analytical techniques. The objective is to identify and model key factors influencing crop productivity, including climatic parameters such as rainfall and temperature, as well as region-specific agricultural practices and input availability. The study demonstrates that while some crops show a strong dependency on weather conditions, others are more significantly affected by local farming methods and resource use.

To improve prediction accuracy, the project integrates advanced Machine Learning (ML) and Deep Learning (DL) approaches alongside traditional models like Linear Regression and Time-Series analysis. These models collectively utilize both historical yield records and external environmental variables to uncover meaningful patterns and trends. The outcomes of this study aim to support data-driven, informed decision-making in the agricultural ecosystem.

2. Introduction

Agriculture is a cornerstone of the Indian economy, providing livelihood to a significant majority of the population. It is intricately linked to national development, food security, and rural employment. The performance of this sector influences not only farmers, but also a wide spectrum of stakeholders including government agencies, private enterprises, and supply chain intermediaries. In such a context, accurately predicting crop yield becomes a critical tool for decision-making in agricultural planning, policy formulation, and resource management.

Agriculture continuously seeks innovative methods to boost productivity and ensure food security. The convergence of data analytics, machine learning, and deep learning has emerged as a transformative force in this pursuit. In the context of India, where agriculture is highly sensitive to climatic and environmental variations, leveraging data-driven techniques for crop yield prediction has become increasingly vital.

Additionally, regional variations in farming practices and irrigation infrastructure introduce further complexity. The datasets used in this study reflect these real-world conditions by capturing daily agro-meteorological variables like maximum and minimum temperatures, rainfall, and water reservoir levels, along with corresponding crop yields across different states and time intervals.

By evaluating the predictive accuracy of these models, the study aims to identify the most effective algorithms and features influencing crop output. Ultimately, this research aspires to enhance the precision of agricultural forecasting and support more informed and timely agricultural decision-making through the integration of machine learning and domain-specific insights.

Topics Covered in the First Two Weeks

During the initial phase of the internship, we were introduced to a diverse set of foundational and technical topics, including:

- Power BI: Basics of data visualization and dashboard creation.
- Streamlit: Create interactive web app.
- Research Project Introduction: Overview of project scope, objectives, and expected outcomes.
- Introduction to Machine Learning and Deep Learning.
- Prompt Engineering & Generative AI: Fundamentals of crafting effective prompts and understanding AI outputs.
- Text Analytics: Introduction to NLP techniques for extracting insights from textual data.

3. Project Objective

- To analyse historical agricultural data to identify key factors influencing crop yield, including temperature, rainfall, and regional practices.
- To develop predictive models using Deep Learning techniques for accurate crop yield forecasting.
- To compare the performance of ANN model and determine the most effective parameter for yield prediction.
- To demonstrate how data-driven forecasting can support strategic planning in agriculture, such as crop selection, resource allocation, and market readiness.

4. Methodology

This project involved a systematic and data-driven approach to analyze and forecast crop yield using historical datasets from various Indian states. The steps followed are as outlined below:

a. Data Collection

- Data was sourced from crop datasets and repositories, including crop yield statistics and agro-meteorological records.
- The primary variables collected included:
  - Crop name and state
  - Daily rainfall and temperature (max/min)
  - Water reservoir levels (FRL, storage)
  - Crop yield for specific intervals (yearly)

b. Data Cleaning

- Removed null values and duplicate entries.
- Converted date fields to standard datetime format.
- Handled inconsistent entries (e.g., wrong units or formatting).
- Ensured data consistency across different states and years.

c. Data Preprocessing

- Feature Engineering: Created new features such as average temperature, growing season indicators, etc.
- Label Encoding for categorical variables (e.g., crop_name, state_name).
- Normalization applied to numeric features to improve model performance.

d. Exploratory Data Analysis (EDA)

- Visualized trends using line plots and bar charts (e.g., yield vs. rainfall).
- Correlation heatmaps to study relationships between variables.
- Outlier detection and handling.
- Identified strong influencing factors for specific crops.

e. Model Selection

- Multiple models were considered:
  - **Artificial Neural Network (ANN)**: Used as the deep learning model for this project. The ANN was designed with an input layer, multiple hidden layers, and an output layer to learn complex non-linear relationships in the data. It was trained using backpropagation and optimized with Adam optimizer and Mean Squared Error (MSE) loss function.
  - **Long Short-Term Memory (LSTM)**: A type of Recurrent Neural Network (RNN) designed to capture long-term dependencies in sequential data. LSTM was particularly effective in modeling temporal trends like rainfall and temperature over time, making it suitable for time-based yield prediction.
- Selection was based on initial performance metrics and explainability.

f. Model Training & Validation

- Dataset was split into **80% training** and **20% testing** sets.
- Cross-validation techniques used to avoid overfitting.
- Hyperparameter tuning performed for models.

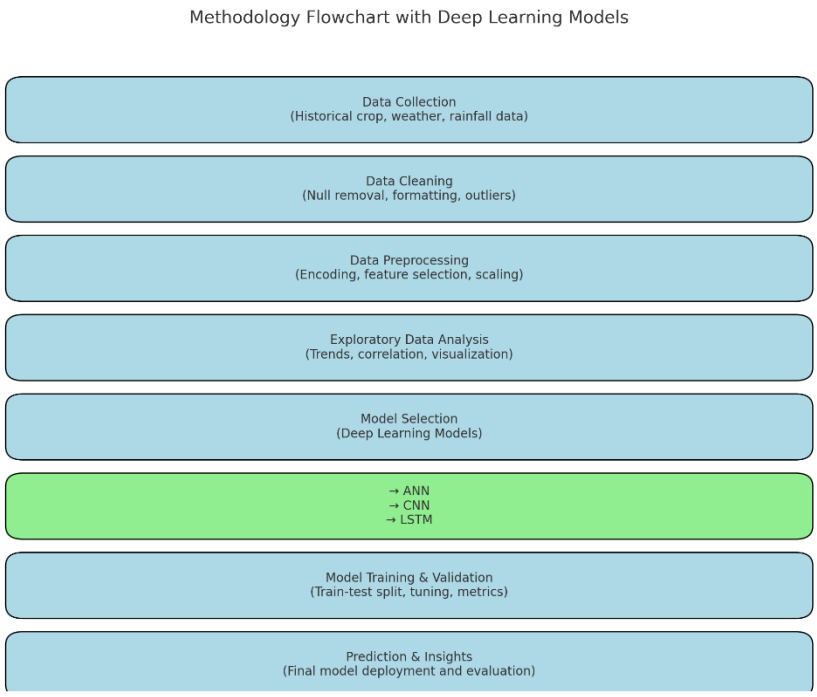
g. Model Evaluation

- Evaluated using:
  - **Root Mean Squared Error (RMSE)**
  - **R^2 Score**
- Compared model performances to select the best forecasting model.

h. Prediction & Insights

- Final model used to predict yields on unseen data.
- Analysis revealed varying dependencies: some crops were rainfall-sensitive while others relied more on temperature or region-specific practices.

These insights were structured to help support data-driven agricultural planning.



Model Performance Evaluation:

During the implementation phase of the deep learning models, we observed notable differences in how well each model generalized on the dataset, particularly in relation to variables such as water reservoir levels.

- LSTM and RNN models, though powerful for capturing sequential and spatial patterns, did not generalize well when applied to this dataset. Specifically, these models struggled to effectively learn from features like water storage levels (FRL, Live Storage, etc.), which are critical yet complex indicators of crop productivity. As a result, both models yielded poor predictive performance, reflected by negative R^2 scores and high RMSE values during testing. This indicates that their predictions were worse than simply predicting the mean yield value.
- In contrast, the Artificial Neural Network (ANN) model showed robust generalization across all key features, including meteorological parameters and water resource indicators. It was able to capture the underlying relationships between rainfall, temperature, water reservoir levels, and crop yield more effectively. The ANN model delivered a higher R^2 score and a lower RMSE, indicating better prediction accuracy. Moreover, the predicted yield values were consistently close to the actual mean yield values for the specific crop and state, which reinforced the ANN's suitability for this task.

5. Data Analysis and Result

The following is the EDA and explanation for the crop dataset:

GRAM

Gram Yield Model Performance (State-wise):

The model's performance was evaluated using RMSE (Root Mean Square Error) and R^2 (Coefficient of Determination) across 14 states:

- The average actual yield across all states was 1.067 t/ha, and the predicted values for each state closely match this, indicating good consistency and generalization.
- RMSE values ranged from 0.116 (Andhra Pradesh) to 0.128 (Madhya Pradesh), showing low prediction errors, with all states under 0.13 RMSE.
- R^2 values ranged between 0.45 and 0.55, indicating that the model explains roughly 45–55% of the variance in yield across different states.
    - The best performance was observed in Andhra Pradesh ($R^2 = 0.550$) and Maharashtra ($R^2 = 0.539$).
    - Relatively lower R^2 values were seen in Madhya Pradesh (0.450) and Karnataka (0.472), suggesting more variability in those regions possibly due to unaccounted local factors.

State	RMSE	R^2 Score	Actual Yield	Predicted Yield
Andhra Pradesh	0.116037	0.550607	1.067169	1.064435
Chhattisgarh	0.123865	0.487932	1.067169	1.068291
Gujarat	0.119840	0.520666	1.067169	1.072419
Jharkhand	0.122651	0.497922	1.067169	1.072787
Karnataka	0.125686	0.472765	1.067169	1.066908
Madhya Pradesh	0.128263	0.450917	1.067169	1.061033
Maharashtra	0.117524	0.539020	1.067169	1.068807
Odisha	0.123329	0.492354	1.067169	1.065131
Rajasthan	0.117943	0.535722	1.067169	1.068408
Tamil Nadu	0.121015	0.511223	1.067169	1.076361
Telangana	0.117691	0.537708	1.067169	1.068578
Uttarakhand	0.120770	0.513202	1.067169	1.068321

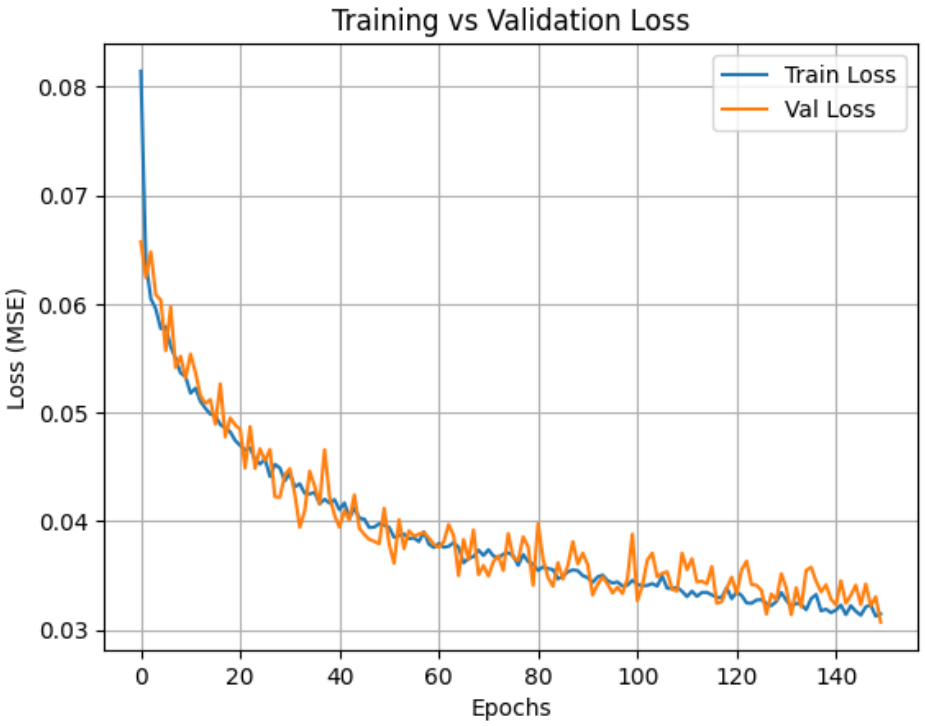
State	RMSE	R^2 Score	Actual Yield	Predicted Yield
Uttar Pradesh	0.119767	0.521251	1.067169	1.059243
West Bengal	0.118809	0.528878	1.067169	1.056500

Table 1: Model Performance Summary for Gram Crop

Model Training Performance:

The plot shows that both training and validation loss steadily decrease over epochs, converging around epoch 100, and continue to decline with minimal gap.

- The low and closely aligned losses indicate the model is learning effectively and generalizing well without overfitting.
- Final loss values stabilize near 0.03 (MSE), suggesting good predictive accuracy.



RICE

Rice Yield Model Performance (State-wise)

The model performance was evaluated using RMSE and R^2 Score for rice yield prediction across five states:

- The actual yield was consistently 3.0677 t/ha, and the predicted yields were very close, showing high prediction accuracy.
- RMSE values are all below 0.101, with Karnataka achieving the lowest error at 0.0957.
- R^2 scores range from 0.644 to 0.680, indicating the model explains around 64–68% of the yield variance in different states.
    - Best performance: Karnataka (R^2 = 0.6797)
    - Slightly lower: Telangana (R^2 = 0.6447

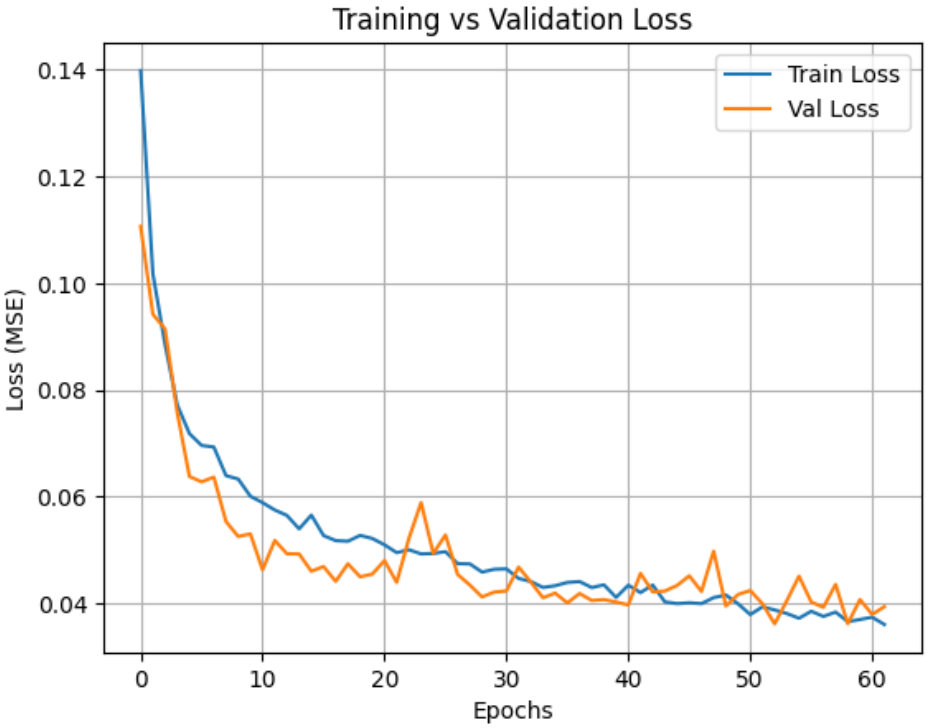
State	RMSE	R^2 Score	Actual Yield	Predicted Yield
Andhra Pradesh	0.097819	0.665548	3.067669	3.068301
Jharkhand	0.100537	0.646703	3.067669	3.072704
Karnataka	0.095728	0.679694	3.067669	3.056760
Telangana	0.100820	0.644712	3.067669	3.078905
Uttarakhand	0.097553	0.667361	3.067669	3.067074

Table 2: Model Performance Summary for Rice Crop

Model Training Performance:

The plot shows the Mean Squared Error (MSE) loss for training and validation datasets over 60 epochs.

- Rapid Initial Decrease: Both losses drop sharply in the first few epochs, indicating quick learning.
- Gradual Convergence & Stability: After initial rapid decline, both losses continue a slower, steady decrease and largely converge, hovering between 0.04 and 0.05 in later epochs. The validation loss shows minor fluctuations but generally tracks the training loss closely.
- Good Generalization: The close proximity and low values of both training and validation loss towards the end suggest that the model is learning effectively and generalizing well to unseen data without significant overfitting.



WHEAT

Wheat Yield Model Performance (State-wise):

The model was evaluated on its ability to predict wheat yield using RMSE and R^2 scores:

- The actual yield is uniform at 2.6217 t/ha.
- The predicted yields range from 2.6198 to 2.6330 t/ha, showing high accuracy and minimal deviation.
- RMSE values are low, between 0.195 and 0.201, indicating small absolute errors in prediction.
- R^2 scores range from 0.584 to 0.608, reflecting the model's ability to explain about 58–61% of the yield variance across states.

- Highest R^2 (0.608): Gujarat — indicating the best model performance in capturing wheat yield variability.
- Lowest R^2 (0.584): Jharkhand — where the model is slightly less explanatory, but still acceptable.

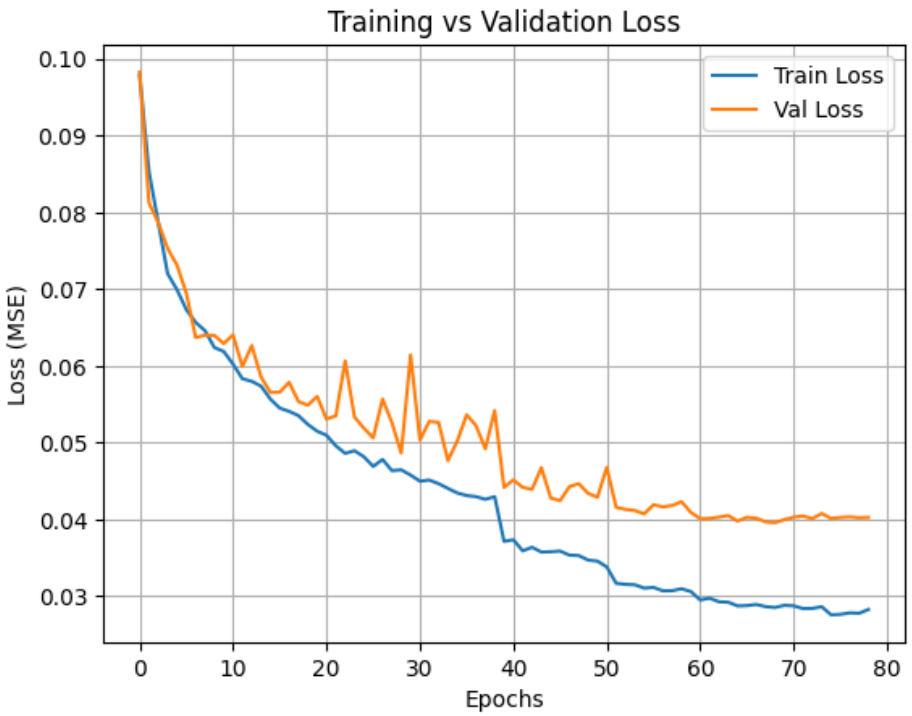
State	RMSE	R^2 Score	Actual Yield	Predicted Yield
Andhra Pradesh	0.204720	0.568657	2.621672	2.628211
Chhattisgarh	0.213630	0.530293	2.621672	2.627347
Gujarat	0.203142	0.575278	2.621672	2.629190

State	RMSE	R^2 Score	Actual Yield	Predicted Yield
Jharkhand	0.204761	0.568483	2.621672	2.630168
Karnataka	0.205576	0.565039	2.621672	2.627720
Madhya Pradesh	0.211378	0.540142	2.621672	2.628400
Maharashtra	0.206081	0.562902	2.621672	2.628684
Odisha	0.208535	0.552428	2.621672	2.622731
Rajasthan	0.215041	0.524068	2.621672	2.631819
Tamil Nadu	0.204572	0.569280	2.621672	2.627793
Telangana	0.203849	0.572317	2.621672	2.627931
Uttarakhand	0.209138	0.549838	2.621672	2.626118
Uttar Pradesh	0.210258	0.545004	2.621672	2.630609
West Bengal	0.202122	0.579532	2.621672	2.625181

Table 3: Model Performance Summary for Wheat Crop

Model Training Performance:

- Effective Initial Learning: Both the training and validation losses start high and exhibit a rapid decrease in the initial epochs, indicating that the model is learning quickly from the data.
- Clear Overfitting Indication: A distinct gap emerges and persists between the training and validation loss curves. The training loss continues to decrease and reaches very low values, while the validation loss plateaus at a higher level (around 0.04 MSE). This divergence is a strong indicator that the model is overfitting to the training data.
- Significant Drop in Training Loss: Around epochs 38-40, the training loss experiences a sharp drop that is not mirrored by the validation loss, further widening the gap and reinforcing the overfitting concern.
- Stable but Diverged Convergence: Both loss curves eventually stabilize towards the end of the training, with the training loss near 0.028 MSE and the validation loss near 0.04 MSE. While stable, the persistent gap indicates that the model's performance on unseen data is notably worse than on the training data.



POTATO

Potato Yield Model Performance (State-wise):

The model was evaluated on its ability to predict potato yield using RMSE and R^2 scores:

- The actual yield across states is 29.59 t/ha, and predicted values are very close, ranging from 29.66 to 29.84 t/ha, showing high prediction accuracy.
- RMSE values are low, ranging from 0.662 (Telangana) to 0.693 (Andhra Pradesh), indicating minimal prediction error.
- R^2 scores fall between 0.542 and 0.582, meaning the model explains 54–58% of the yield variance across states.
- Best performance: Telangana with lowest RMSE (0.662) and highest R^2 (0.582)
- Slightly weaker performance: Andhra Pradesh with highest RMSE (0.693) and lowest R^2 (0.542).

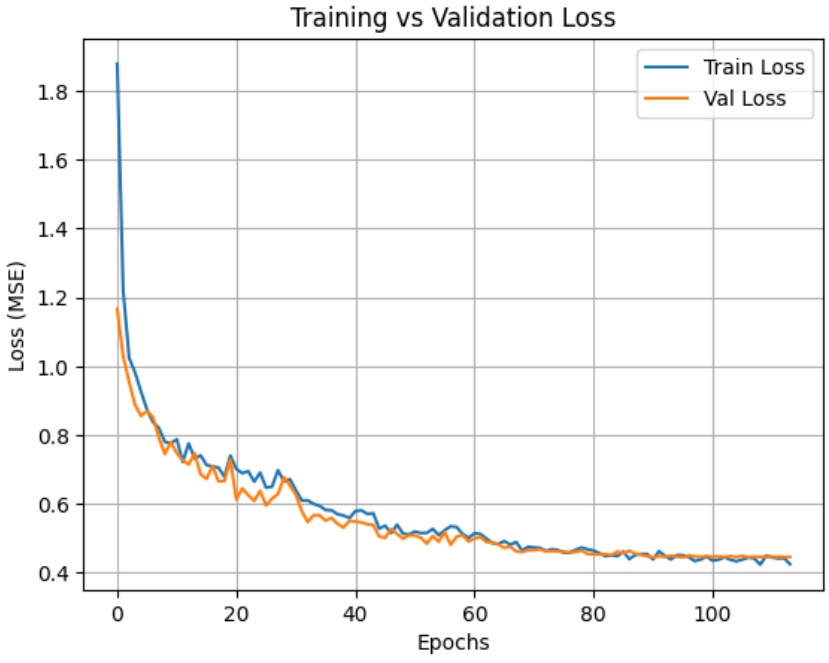
State	RMSE	R^2 Score	Actual Yield	Predicted Yield
Andhra Pradesh	0.693650	0.542268	29.593285	29.835829
Chhattisgarh	0.669063	0.574142	29.593285	29.836721
Jharkhand	0.674199	0.567579	29.593285	29.816730
Karnataka	0.680787	0.559087	29.593285	29.835361

State	RMSE	R^2 Score	Actual Yield	Predicted Yield
Tamil Nadu	0.683481	0.555591	29.593285	29.792906
Telangana	0.662502	0.582454	29.593285	29.737795
Uttarakhand	0.664586	0.579822	29.593285	29.660210
Uttar Pradesh	0.678009	0.562678	29.593285	29.801168
West Bengal	0.666716	0.577126	29.593285	29.769056

Table 4: Model Performance Summary for Potato Crop

Model Training Performance:

- High Initial Loss & Rapid Drop: Both training and validation losses begin at a very high level (around 1.9 and 1.2 MSE, respectively) but experience a sharp and significant decrease in the first 20-30 epochs, indicating effective initial learning.
- Strong Convergence: After the rapid initial drop, the training and validation loss curves quickly converge and remain very close to each other throughout the remainder of the training process (from approximately epoch 30 onwards).
- Stable & Consistent Performance: Both losses stabilize and show minimal fluctuations in the later epochs (from around epoch 50 to 110+), indicating a consistent and robust learning process.
- Good Generalization: The close alignment and stable nature of the training and validation loss curves demonstrate that the model is generalizing very well to unseen data without significant signs of overfitting.
- Final Loss Values: The final MSE values for both training and validation loss stabilize around 0.43 to 0.45. While the generalization is good, the absolute loss value is relatively higher compared to ideal scenarios, suggesting that the model's predictions still have a notable average squared error.



MASSOR

Massor Yield Model Performance (State-wise):

The model's performance was evaluated across 9 states using RMSE and R^2 Score:

- The actual yield across all states is fixed at 0.8055 t/ha.
- Predicted yields range from 0.8025 to 0.8079 t/ha, showing high numerical closeness to the actual value.
- RMSE values are very low, between 0.104 and 0.108, reflecting minimal absolute prediction error.
- R^2 scores range from 0.343 to 0.391, meaning the model explains only about 34–39% of the variance in yield across regions.

- The model delivers accurate predictions near the mean, as shown by low RMSE**.**
- However, the moderate R^2 values indicate that while the model captures the overall trend, it struggles with regional variability.

State	RMSE	R^2 Score	Actual Yield	Predicted Yield
Chhattisgarh	0.104995	0.383485	0.805541	0.805779
Jharkhand	0.107635	0.352091	0.805541	0.802902
Madhya Pradesh	0.107176	0.357604	0.805541	0.802804
Odisha	0.108404	0.342796	0.805541	0.802484
Rajasthan	0.105860	0.373279	0.805541	0.806166
Telangana	0.106178	0.369514	0.805541	0.807542
Uttarakhand	0.106670	0.363657	0.805541	0.804645
Uttar Pradesh	0.104352	0.391011	0.805541	0.807903
West Bengal	0.107457	0.354228	0.805541	0.802568

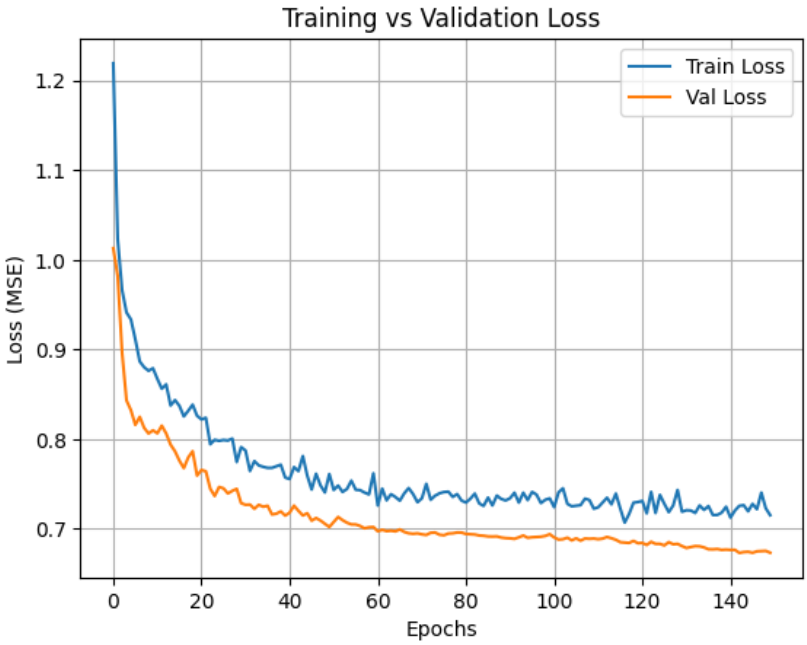
Table 5: Model Performance Summary for Massor Crop

Model Training Performance:

- High Initial Loss and Rapid Descent: Both training and validation losses start at very high values (over 1.2 for training and 1.0 for validation) and exhibit a sharp, rapid decrease in the initial epochs, indicating effective early learning.
- Unusual Loss Relationship: After the initial drop, the validation loss consistently drops below the training loss and remains lower for the majority of the training process. This is an unusual pattern, suggesting that the model performs better on the

unseen validation data than on the training data itself. This could be due to strong regularization techniques applied during training (e.g., dropout, L1/L2 regularization penalizing the training loss more), or potentially a less complex validation dataset.

- Stabilization at Moderate Levels: Both loss curves eventually stabilize, with the validation loss settling around 0.67 MSE and the training loss fluctuating around 0.72-0.74 MSE in the later epochs.
- Overall Stability: The model shows stable convergence, as both losses flatten out towards the end of the training process, indicating that further training would likely yield minimal improvements



MUSTARD

Mustard Yield Model Performance (State-wise):

The model was tested across 14 states for predicting mustard yield:

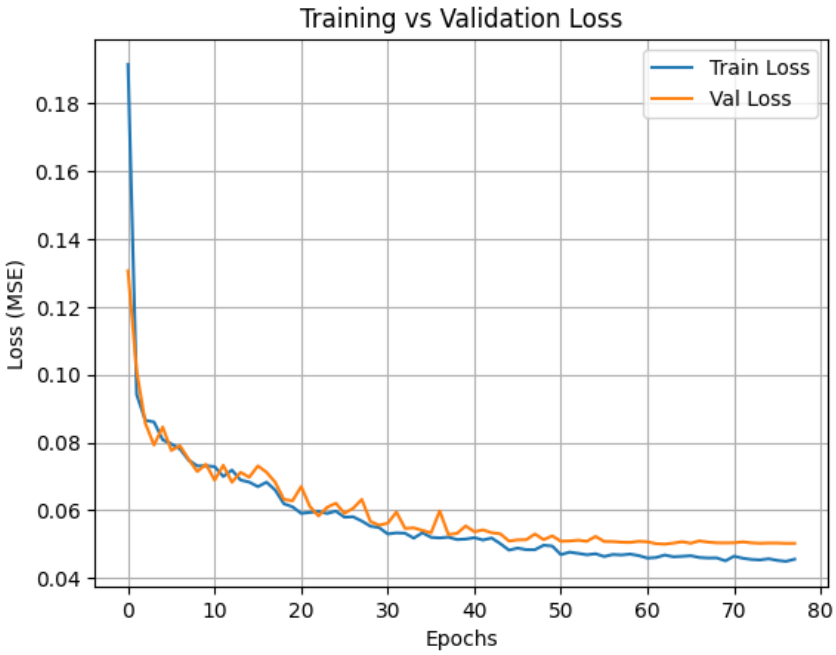
- The actual yield is constant at 1.0051 t/ha across all states.
- The predicted yields range from 1.0031 to 1.0123 t/ha, showing very high closeness to the actual value.
- RMSE values are low, ranging from 0.1005 to 0.1102, indicating small prediction errors.
- R^2 scores fall between 0.525 and 0.605, meaning the model explains approximately 52–60% of the variance in yield data.
- Low RMSE (~0.1) confirms the model's accuracy in numeric prediction.
- The moderate R^2 values suggest the model balances both accuracy and generalization well.

State	RMSE	R^2 Score	Actual Yield	Predicted Yield
Andhra Pradesh	0.101283	0.599801	1.005127	1.008512
Chhattisgarh	0.101428	0.598657	1.005127	1.007828
Gujarat	0.100532	0.605717	1.005127	1.006918
Jharkhand	0.103643	0.580935	1.005127	1.010761
Karnataka	0.107010	0.553264	1.005127	1.010048
Madhya Pradesh	0.101867	0.595172	1.005127	1.008000
Maharashtra	0.103451	0.582483	1.005127	1.010151
Odisha	0.110256	0.525750	1.005127	1.003147
Rajasthan	0.101753	0.596080	1.005127	1.007936
Tamil Nadu	0.101185	0.600578	1.005127	1.009654
Telangana	0.109852	0.529218	1.005127	1.012311
Uttarakhand	0.102576	0.589514	1.005127	1.007313
Uttar Pradesh	0.102231	0.592277	1.005127	1.006058
West Bengal	0.104608	0.573096	1.005127	1.010681

Table 6: Model Performance Summary for Mustard Crop

Model Training Performance:

- Effective Initial Learning: Both training and validation losses start high but exhibit a rapid and significant decrease within the first 5-10 epochs, demonstrating efficient initial learning by the model.
- Good Generalization with Minor Overfitting: After the initial phase, the validation loss consistently remains slightly above the training loss, indicating a minor degree of overfitting. However, the gap is small and stable, suggesting the model is still generalizing well to unseen data.
- Stable Convergence: Both loss curves show a clear trend towards stabilization from approximately epoch 40-50 onwards, signifying that the model has largely converged and further training would yield diminishing returns.
- Low Final Loss Values: The final Mean Squared Error (MSE) values for both training (around 0.045) and validation (around 0.05) are relatively low, suggesting that the model has achieved good predictive accuracy for the mustard crop.



6. Conclusion

This study employed regression-based approaches, particularly artificial neural networks (ANNs), to forecast crop yields using historical data combined with auxiliary variables such as temperature, rainfall, and reservoir levels. The inclusion of these auxiliary variables, carefully selected based on crop growth phases and agro-climatic relevance, significantly contributed to the model's predictive performance.

Across various crop including gram, rice, wheat, potato, mustard, massor. The models consistently achieved low RMSE values, indicating high accuracy in yield estimation. While R^2 scores varied moderately across regions and crops, they reflected the model's reasonable capability in capturing the variance in yield patterns. Notably, in high-yield crops like potato and rice, the relative error remained low**,** confirming the reliability of the models for practical applications.

The results demonstrate that such predictive models can support timely decision-making for both farmers and policymakers. Farmers can use the yield forecasts to optimize farm practices, while governments can utilize the insights for strategic planning and resource allocation. Unlike traditional linear models, the ANN employed in this study leverages nonlinear transformations, enabling it to model complex relationships between crop yield and climatic/auxiliary variables more effectively.

7. Appendices

Dharmaraja, S., Jain, V., Anjoy, P. *et al.* Empirical Analysis for Crop Yield Forecasting in India. *Agric Res* **9**, 132–138 (2020). https://doi.org/10.1007/s40003-019-00413-x

Ujjainia*, S., Gautam, P., & Veenadhari, S. (2020). Crop Yield Prediction using Regression Model. International Journal of Innovative Technology and Exploring Engineering, 9(10), 269–273. https://doi.org/10.35940/ijitee.j7491.0891020

Collab File Link-
https://colab.research.google.com/drive/1QEqBkW4et6ODkaC9qijdMDDusuVoye2A?usp=sharing

Github Link-

https://github.com/Navneetsingh9/ISI-CROP-YIELD-ANALYSIS-AND-FORECASTING