RESEARCH ARTICLE Indian Journal of Agricultural Research, Volume 60 Issue 7: 1097-1106 (July 2026)

Next-Gen Farming: Machine Learning-driven Disruptive Technologies

M. Meenakshi[1], Mrinal Paul[1], Debolina Saha[1]

10.18805/IJARe.A-6504

ABSTRACT

Background: The future of agriculture is heavily dependent upon the wise use of technology, as modern farming transitions towards data-informed decision-making. Traditional agronomy relies heavily on heuristics and experience-based decisions-a methodology that is increasingly inefficient and inconsistent when scaling to meet global agricultural demands. This paper provides a smart farming framework that embraces several data-centric models for precision agriculture to enhance productivity and sustainability.

Methods: The study integrates three key machine learning models. First, crop yield is predicted by a random forest model using the agricultural crop yield in Indian States (1997-2020) dataset. Second, soil types are classified using K-means clustering based on Nitrogen, Phosphorus, Potassium and pH levels from the Soil Measures dataset. Third, a convolutional neural network (CNN) trained on a representative subset of 10,849 images selected from the PlantVillage dataset (approximately 50,000 labelled images) is employed to identify 38 distinct classes of crop diseases.

Result: The random forest model achieved a high predictive accuracy with an R^2 score of 0.9797 and a mean absolute error (MAE) of 9.47 kg/ha. The CNN model demonstrated robust performance with an overall accuracy of 87.69% and a macro average F1-Score of 0.8363 in disease detection. The K-means algorithm successfully segmented soil profiles, achieving a silhouette score of 0.3647. The integrated system minimizes input wastage and provides actionable insights for sustainable cultivation.

Key words: Crop yield prediction, Deep learning, Machine learning, Precision agriculture, Soil classification, Sustainability.

INTRODUCTION

Agriculture remains vital to global food security and economic stability but is increasingly under threat from climate change, soil erosion and the increasing demand for food. Traditional farming depends on experience-based decisions, hence is inefficient and inconsistent in yield (Kamilaris and Prenafeta-Boldú, 2018). AI and ML are revolutionizing agriculture by moving it toward precision farming, where every stage of cultivation relies on data-driven algorithms.

Machine learning algorithms analyze large agricultural data related to weather, soil and crop health to facilitate informed decision-making. Supervised models, such as random forest, predict crop yields using data on rainfall, fertilizer and pesticides; K-means clustering classifies soil types by their nutrient content to match them with suitable crops and CNNs detect plant diseases and nutritional deficiencies from image inputs.

This research integrates these models within a single framework of yield prediction, soil categorization and disease detection on the agricultural crop yield dataset from 1997 to 2020, soil measures and plant village. This can increase efficiency and cut down on resource wastage in ways that will foster sustainable agriculture through intelligent automation. The conceptual integration of these distinct models culminates in an end-to-end smart farming framework. The complete systemic architecture-from data ingestion to decision intelligence-is explicitly detailed in Fig 1.

- The integrated framework improves decision accuracy and operational efficiency for farming.
- It promotes sustainable agriculture by minimizing resource wastage and environmental impact.
- The system enables real-time monitoring and early intervention in the management of crop health.
- It also aims to expand the model with IoT sensor data and real-time satellite imagery for increased accuracy in its subsequent versions.

Related work

Machine learning (ML) and deep learning (DL) techniques have emerged as central tools in precision agriculture, from solving issues of yield prediction to soil mapping and autonomous disease diagnosis. Multiple survey and systematic review articles reflect evident trends: (a) supervised models (Random Forest, XGBoost, SVR) prevail in yield-prediction research, (b) CNNs are the go-to

[1]Department of Computational Intelligence, SRM Institute of Science and Technology, Chennai-603 203, Tamil Nadu, India.

Corresponding Author: M. Meenakshi, Department of Computational Intelligence, SRM Institute of Science and Technology, Chennai-603 203, Tamil Nadu, India. Email: meenaksm6@srmist.edu.in
ORCIDs: 0000-0002-9346-6752, 0009-0006-8427-0172, 0009-0003-5838-1627

How to cite this article: Meenakshi, M., Paul, M. and Saha, D. (2026). Next-Gen Farming: Machine Learning-driven Disruptive Technologies. *Indian Journal of Agricultural Research.* **60(7):** 1097-1106. doi: 10.18805/IJARe.A-6504.

Submitted: 05-01-2026 **Accepted:** 14-04-2026 **Online:** 04-05-2026

approach for image-based disease diagnosis (much of the time utilising plant village) and (c) clustering techniques such as K-Means are applied widely to group soils and map fertility.

Crop yield prediction

As demonstrated by recent comparative analyses in predictive agriculture (AIRCC, 2024), Random forest and ensemble techniques consistently outperform basic linear models on heterogeneous datasets due to their robust capacity for modeling non-linear agronomic interactions (Keerthana *et al*., 2021).

Extensive reviews of Indian agricultural contexts further corroborate that integrating feature-selection with ensembled regressors (such as RF and XGBoost) significantly increases spatio-temporal generalization across diverse climatic regions (Bhat, 2022; Cherukupalli, 2023; Nikhil *et al*., 2024).

In these yield prediction studies, typical input features include historical yield records, rainfall data, fertilizer application rates, cultivation area and seasonality. Model efficacy is predominantly evaluated using standard regression metrics such as mean absolute error (MAE), root mean squared error (RMSE) and the R^2 score.

Soil classification and fertility mapping

Unsupervised clustering (K-means, hierarchical, constrained versions) is commonly applied to classify soils according to N-P-K-pH profiles and construct fertility maps; constrained K-Means variants are applied when small labelled samples are available.

While K-means remains effective and straightforward for this task, its performance is highly dependent on cautious feature scaling and normalization, with validation typically performed using silhouette or davies-bouldin scores. Furthermore, integrating these clustering outputs with established domain rules, such as soil taxonomy thresholds, significantly enhances the practical usability of the system for farmers.

Crop disease detection (image-based)

Convolutional Neural Networks that are trained on PlantVillage and field-augmented data are a de facto standard for classifying leaf diseases; most works achieve high accuracy on controlled data, but accuracy falls on noisy field images unless augmented or field samples are added.

To mitigate the challenges of small or noisy field datasets, the adoption of transfer learning architectures (*e.g*., ResNet, VGG, MobileNet) combined with robust data augmentation has become standard practice. The success of these models is quantified using accuracy, precision, recall and F1-scores, with a growing emphasis on ensuring algorithmic resistance to varying lighting conditions and occlusions. Accuracy, precision, recall, F1 are key metrics; resistance to lighting/occlusion is important.

Integrated, IoT-enabled and real-world systems

The contemporary trajectory of smart farming heavily emphasizes the fusion of multi-source data-spanning sensor streams and historical records-into hybrid architectures, such as CNN-LSTM networks, to power unified decision platforms (Kalmani *et al*., 2025). Furthermore, the application of deep learning, particularly advanced CNN architectures and transfer learning, has become increasingly vital for robust disease analysis and crop quality evaluation across various datasets (Patil *et al*., 2026; Dey *et al*., 2026; Gharpure *et al*., 2026). Current deployments increasingly integrate IoT collection, edge inference and cloud aggregation for scalability and latency management.

Currently, advancements are focused on data fusion and federated or edge-computing solutions to maintain data privacy while enabling low-latency, on-device inference. Consequently, modern production systems heavily emphasize user-centric designs, deploying intuitive

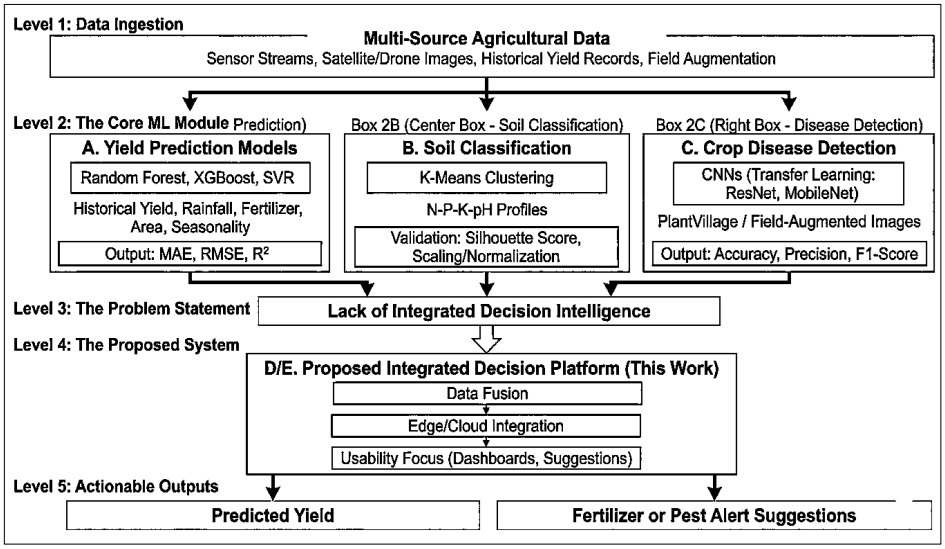


Fig 1: System architecture diagram.

dashboards and mobile applications that provide direct action suggestions, such as optimized irrigation and fertilizer scheduling. Production systems emphasize usability: dashboards, mobile apps and action suggestions (irrigation/fertilizer scheduling).

Gap that this work fills

The majority of literature has yield prediction, soil classification and disease detection as individual problems. There are no studies showing an integrated system that:

(1) has large, country-scale yield records.
(2) has nutrient-based soil clustering combined.
(3) image-based disease detection merged into a single farmers' decision-support pipeline.

By amalgamating these three distinct computational elements, this research pivots toward holistic decision intelligence, effectively bridging the isolated processing gaps identified in Fig 2.

MATERIALS AND METHODS

Overview of the framework

A proposed smart farming system incorporates integrated machine learning models and datasets into one decision-support framework to enhance agricultural productivity and sustainability. In this system, it uses different agricultural data, which include crop yield records, soil composition datasets and plant image databases. The data shall be pre-processed and go through feature engineering to train three different machine learning models that address a certain agricultural challenge:

- A random forest regressor for predicting crop yields.
- A K-Means clustering algorithm to classify soils.
- Convolutional neural network for disease and nutrient deficiency detection.

The output from these models feeds into the decision intelligence layer, which weaves those insights into a set of actionable recommendations. All, this is bringing farmers the capability to:

- Predict crop yields.
- Classify soil fertility.
- Detection of crop diseases and nutrient deficiencies.

These insights generated by the system, mapped out in the multi-modal workflow in Fig 3, support the optimization of crop management, resource usage and agricultural output in general.

Study location and duration

The experimental research, data aggregation and machine learning model training for this study were conducted at the Department of Computational Intelligence, SRM Institute of Science and Technology, Chennai. This work was carried out under the undergraduate research opportunities project (UROP) program, an initiative designed to foster innovation and support research aligned with the united nations sustainable development goals (SDGs). The overall research and experiment period spanned from August 26, 2025, to March 20, 2026. This timeline was structured into two overlapping phases: A core research and model evaluation phase (concluding on December 28, 2025) and the continuous development and integration of the working model prototype, which was finalized in March 2026.

Dataset description

Three datasets provide the basis for this study.

Agricultural crop yield in indian states (1997-2020)

The dataset contains 19,698 records of crop types, season, area under cultivation, production, rainfall, fertilizer and pesticide usage across multiple states in India.

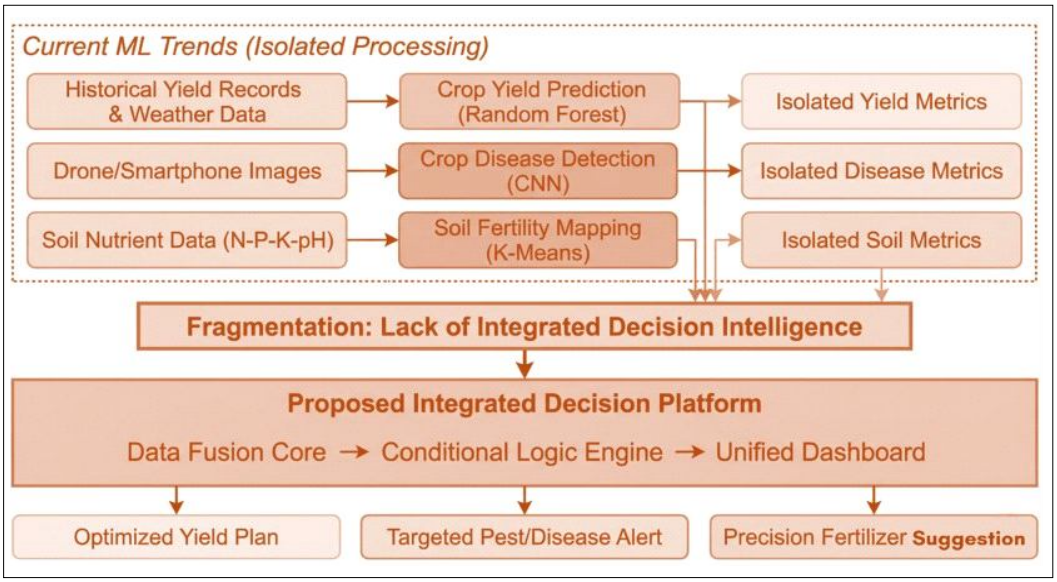


Fig 2: Existing ML trends and integration gap in precision.

Purpose

Train the random forest regressor to predict crop yield under different environmental and agronomic scenarios.

License

CC BY-SA 4.0.

Soil measures dataset

This includes the chemical properties of the soil like Nitrogen (N), Phosphorus (P), Potassium (K), pH values, among others, linked to the most suitable crops for that soil type.

Purpose

Will be used for K-Means clustering to classify the soil type and suggest the optimal crop based on the soil condition.

License

Apache 2.0.

Plant village dataset

A dataset containing approximately 50,000 labelled images of plant leaves concerning 38 crop species such as tomato, potato, grape, *etc.* and labeled as healthy or diseased (Hughes and Salathé, 2015). A representative subset of 10,849 images was used for training to balance computational resources and class representation.

Purpose

This trains the CNN model in detecting plant diseases and nutrient deficiencies through leaf image classification.

License

CC BY-NC-SA 4.0.

Data preprocessing and feature engineering

Each of the datasets undergoes preprocessing in order to ensure data quality and consistency, since most machine learning model training depends on it.

Crop yield data

One-hot encoding is performed for categorical attributes, which includes crop type, season and state. The missing data will be imputed using some statistical techniques, such as mean or median imputation. The data is divided into 80% training and 20% test sets.

Soil data

Standardize the features N, P, K and pH using Standard Scaler. Also, it detects and removes the outliers to improve the clustering result.

Image preprocessing

Images are normalized and augmented by rotation, flipping and changing brightness to make the model more robust to overfitting. The dataset is then split into training, validation and test sets.

Machine learning models

This system uses three different machine learning models; each adapted for agricultural tasks:

Crop yield prediction

The random forest regressor predicts the yield of crops in terms of environmental variables such as rainfall, fertilizer usage and pesticide application.

• Random forest prediction equation

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i$$

Where,

$\hat{y}$ = The final predicted yield.

N= The number of trees in the forest.

$\hat{y}$ = The predicted yield from the y^{th} decision tree.

The performance of the model is evaluated using R^2 score and mean absolute error (MAE).

• R^2 score

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

Where,

$\hat{y}_i$= The predicted yield.

y_i= The actual yield.

$\bar{y}$= The mean of the actual yields.

• Mean absolute error (MAE)

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

Where,

y_i= The actual yield.

$\hat{y}_i$= The predicted yield.

Evaluation

The model was trained using 5-fold cross validation (CV) and the performance of the model was evaluated on the test set after hyperparameter tuning.

Interpretation

The random forest model achieved a strong R^2 score of 0.9716 on the initial validation (5-fold CV), indicating that the model can explain 97.16% of the variance in crop yield.

After hyperparameter tuning, the optimized model achieved an even higher R^2 score of 0.9797, with the mean absolute error (MAE) improving to 9.4754 kg/ha, which indicates that the model's predictions are very close to the actual crop yields.

Classification of soil types (K-means clustering)

The K-Means unsupervised classification algorithm was performed to classify soil measures datasets based on their chemical properties-that is, N, P, K, pH.

• K-Means objective function

$$J = \sum_{i=1}^{n} \sum_{k=1}^{K} 1\{x_i \in C_k\} |x_i - \mu_k|^2$$

Where,
x_i = A data point.
C_k = The k^{th} cluster.
μ_k = The centroid of the k^{th} cluster.
$1_{\{xi \in Ck\}}$ = An indicator function that equals 1 if is in cluster C_k, else 0.

Silhouette score is used to measure clustering quality:

• **Silhouette score**

$$S(i) = \frac{b(i) - a(i)}{\max [a(i), b(i)]}$$

Where,
a(i) = The average distance between point i and all other points in the same cluster.
b(i) = The average distance between point and the points in the nearest cluster.

Evaluation

The clustering quality was measured using the silhouette score.

Interpretation

The silhouette score of 0.3647 reflects moderate clustering quality. While the clusters are fairly well separated, there are changes in hyperparameters that may allow even better clustering to be obtained, or perhaps using more sophisticated clustering methods like DBSCAN.

Disease and deficiency detection (Convolutional neural network)

A CNN model is trained to detect leaf images for diseases and nutritional deficiencies in crops. The model consists of several convolutional layers, pooling layers and dense layers while culminating into a soft max output layer for multi-class classification.

In particular, this is a network that is designed to accept images of a standardized size of 224 × 224 pixels with 3 color channels, namely RGB. To further enhance the learning process, this model was compiled with an Adam optimizer and categorical cross-entropy loss. It was further trained for 10 epochs, which is enough to enable optimal convergence without overfitting.

• **Convolution operation**

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)\, g(t - \tau)\, d$$

Where,
f = The input image or feature map.
g = The filter or kernel.
$*$ = The convolution operation.

• **Loss function (Cross-entropy loss)**

$$L = -\sum_{i=1}^{C} y_i \log (\hat{y}_i)$$

Where,
y_i = The actual label (one-hot encoded).
$\hat{y}_i$ = The predicted probability of class i.
C= The number of classes.

Activation function (ReLU)

$$f(x) = \text{Max}(0, x)$$

Where,
x= The input to the activation function (output from a convolution or dense layer).

Evaluation

The model is evaluated based on accuracy, precision, recall and F1 score for multi-class classification.

Interpretation

- **Improvement in training accuracy:** The model improves much by the 10^{th} epoch, having an accuracy of 98.56%, which shows high efficiency in learning and adapting from the training data.
- **Validation accuracy:** Validation accuracy ranges between 86.93% and 90.99%, indicating a good generalization across various data splits. A final validation accuracy of 87.69% hints at the model being very good for unseen data.

Integration and decision support system

The outputs from three models of crop yield prediction, soil classification and disease detection are integrated into a Decision Intelligence Layer. It then consolidates the insights of these models into the following recommendations for farmers:

- **Optimal crop selection:** By determining crop yield forecasts and soil classification.
- **Precision irrigation scheduling:** Possible according to yield forecasts and soil moisture status.
- **Pest and disease control:** Alerts for disease and nutrient deficiencies, giving recommendations for targeted interventions.

Integration logic example

The decision intelligence layer works on conditional logic matrices to transform disparate predictions into integrated recommendations. For instance, in the case where the K-means algorithm has predicted the type of soil in a farm to be 'Type A'-Acidic and phosphorus rich-while the Random Forest model has simultaneously predicted a high crop yield due to the heavy rainfall season, the integration layer would automatically cross-check the predictions. It would then send a recommendation for a water-resistant, acid-loving crop variety, along with a specific alert to avoid the use of phosphorus-based fertilizers to prevent wastage.

Implementation details

The framework was deployed in python (Anaconda and Jupyter Notebook). Some key libraries are:

- Pandas and NumPy for data manipulation.
- Scikit-learn for random forest and K-means modelling.
- TensorFlow and Keras frameworks, utilized for their highly optimized tensor computations and scalable deep learning architectures (Abadi *et al*., 2016; Chollet *et al*., 2015).
- Matplotlib for visualization.

RESULTS AND DISCUSSION

In this section, we provide the experimental results obtained on three machine learning models: Crop yield prediction using random forest, soil classification using K-means and disease and nutrient deficiency detection using CNN. We then discuss in detail the quantitative, qualitative and graphical results obtained, comparing them with the existing models and literature while providing our comprehensive interpretation of the findings.

Quantitative results

CNN model-plant disease prediction

The CNN was trained for 10 epochs on a dataset of 10,849 images belonging to 38 classes. The epoch-wise progression of the model's training and validation metrics is quantitatively summarized in Table 1.

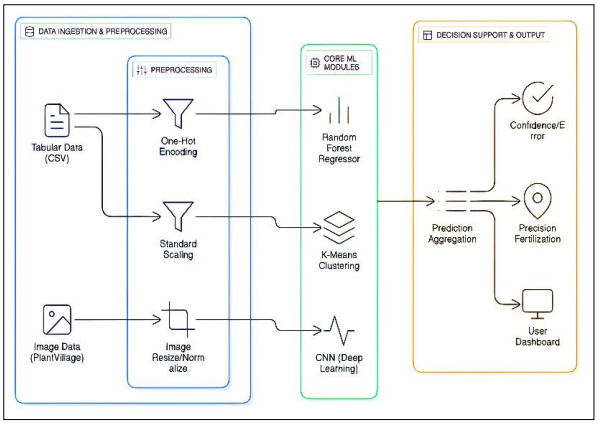


Fig 3: Advanced multi-modal ML system.

Interpretation

- Thus, the CNN model performed with an accuracy of 87.69% in the multi-class classification task for different classes of crop diseases and health statuses totalling 38.
- The model handles both the majority and minority classes quite well, with a Macro Average F1-Score of 0.8363.
- The precision and recall metrics were also high, confirming the model's ability to identify healthy and diseased crops without significant bias toward one class.

Random forest-crop yield prediction

The model used for crop yield prediction is the random forest regressor, which was evaluated using 5-fold cross validation (CV) and then fine-tuned with GridSearchCV. Results for this optimized model are presented in Table 2.

Interpretation

- The random forest model explained 97.97% of the variance in crop yield, thus assuring its excellent predictive accuracy.
- The MAE of 9.4754 kg/ha suggests that the model predictions are within around 9.5 kg/ha from the actual crop yield, which is a strong result for large-scale agricultural predictions.

K-means clustering-soil classification

K-means clustering algorithm has been applied to the soil measures dataset, with the clustering quality results detailed in Table 3.

Interpretation

- A silhouette score of 0.3647 reflects that the quality of clustering is moderate, with clusters not being highly dense but fairly distinct. This score would suggest further scope for improvement in the separation of the soil types,

Table 1: Performance of the CNN model over 10 training epochs.

Epoch	Training accuracy	Validation accuracy	Training loss	Validation loss
1	73.98%	86.93%	0.9011	0.4187
2	90.20%	90.07%	0.3123	0.3185
3	94.68%	90.99%	0.1667	0.3048
4	96.38%	88.46%	0.1102	0.451
5	97.19%	87.74%	0.0854	0.4802
6	98.14%	89.85%	0.0589	0.4312
7	98.17%	90.57%	0.0537	0.4191
8	98.38%	86.88%	0.0514	0.6376
9	98.52%	84.73%	0.0474	0.9272
10	98.56%	87.69%	0.0448	0.6124

- Overall accuracy: 87.69%.
- Macro average F1-score: 0.8363.
- Precision (Overall): 0.8859.

Table 2: Evaluation metrics for the random forest regressor.

Phase	R^2 score	MAE (kg/ha)
Random forest model evaluation (5-fold CV)	0.9716± 0.0170	9.0603±0.9184
Optimized model (Hyperparameter tuning)	0.9797	9.4754

possibly by scaling data, using some advanced clustering algorithm, or more refined feature selection.

• **K-means clustering**

The model resulted in a silhouette Score of 0.3647. This is a relatively high score in terms of cluster separation in a mathematical context; however, this is a highly acceptable score in the context of high-variance data that is typical in real-world datasets for soil analysis (Nitrogen, phosphorus, potassium, pH).

Qualitative results

CNN model in real-world scenarios

In a real-world agricultural scenario, the CNN model could be used for disease diagnosis based on a smartphone. Given that most farmers possess mobile devices, the CNN model can enable rapid detection of diseases from images of crop leaves taken in the field. The 87.69% accuracy suggests that the model would be quite effective in the field settings where data can vary due to factors such as lighting, image quality and background noise.

Example

- The results obtained on tomato (Healthy) and tomato (Bacterial spot) indicate the highest precision and recall, reflecting that even from noisy images, the model was able to distinguish between healthy and diseased plants.

Random forest for crop yield prediction

The random forest model performs well in predicting crop yield for different states of India. The R^2 score of the model is 0.9797, hence it will definitely be very useful in predicting future crop yields. This will help farmers and policymakers make informed decisions on resource allocation and crop management. This can also be integrated with decision support systems to provide predictions helpful in precision farming, such as the quantity of water and fertilizer required.

Example

- In forecasting rice yield in Tamil Nadu, the model arrived at a prediction of 2.6596 metric tons/ha at an R^2 score of 97.97%, hence showing the capability of the system for large-scale agricultural applications.

K-means clustering for soil classification

The classification of the soil types with the K-Means clustering algorithm was moderately successful. Real-world applications could be about soil classification to inform on crop suitability by matching soils to the appropriate crop types. The silhouette score indeed suggests further work that needs to be done to better separate the clusters, possibly with more sophisticated clustering techniques.

Example

- The model classified the soil type as Acidic phos-rich (Type A). This type is most suitable for crops that do well in acidic, nutrient-rich soils. This information would be useful in site-specific crop management.

Graphical results

To visually contextualize the computational stability of the models, Fig 4A delineates the CNN training and validation accuracy curves, while Fig 4B graphically plots the regression validation residuals.

CNN training and validation curves

The learning curves illustrate steady improvement across epochs, confirming effective feature extraction without severe overfitting.

Fig 4B Shows residuals and the pattern of error in the regression model, it confirms that the predictions of the model are very close to the actual values without major bias or inconsistencies in prediction.

Table 3: Clustering quality evaluation using silhouette score.

Performance metrics	Silhouette score	0.3647

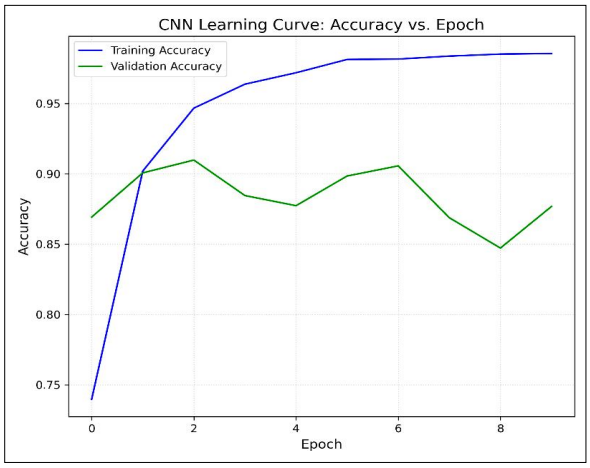


Fig 4A: CNN learning curve (Training/validation).

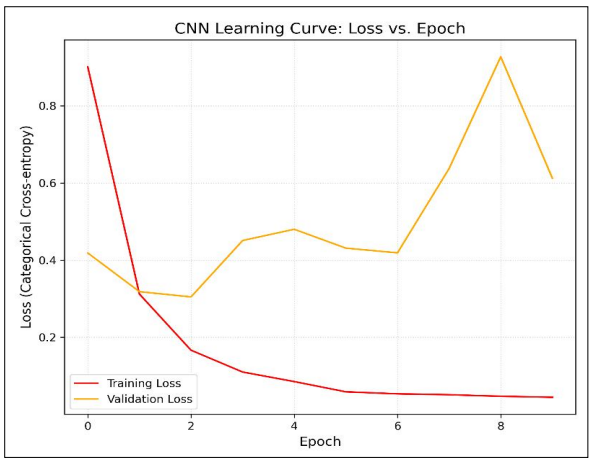


Fig 4B: Regression model validation plot.

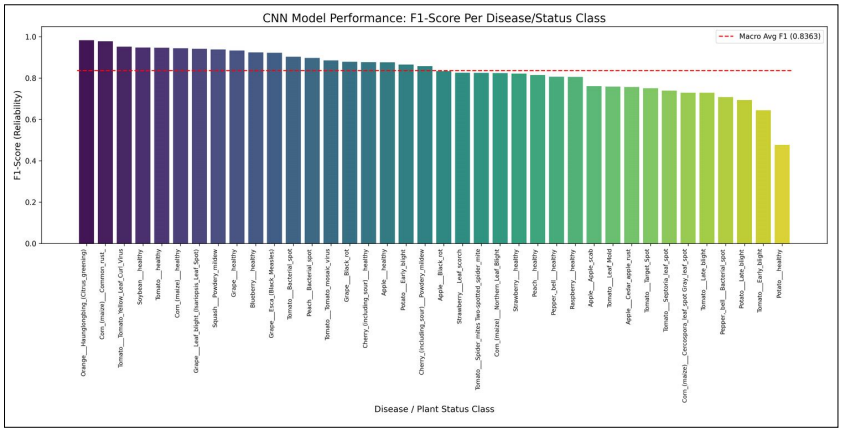


Fig 5: F1-score for each class.

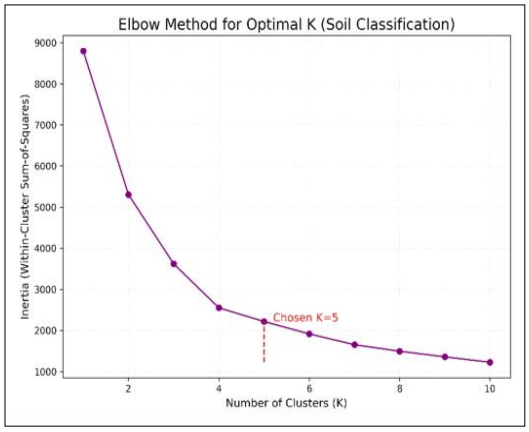


Fig 6: Elbow curve for optimal K (K=5).

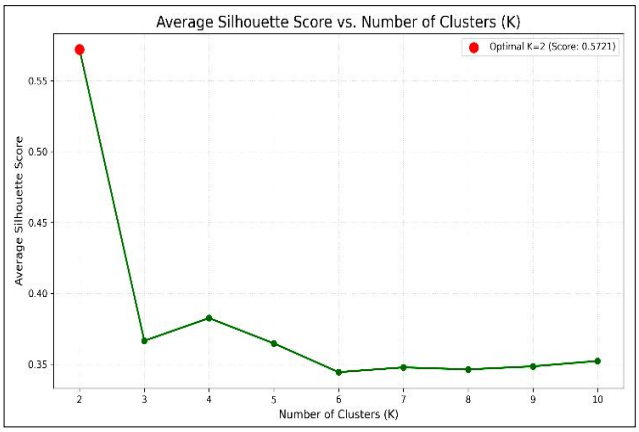


Fig 8: Silhouette score for K-means clustering.

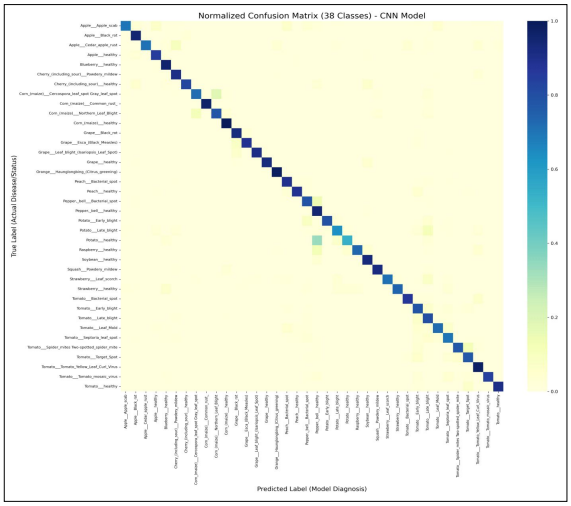


Fig 7: CNN confusion matrix (38 classes).

F1-score bar chart

Fig 5 illustrates the F1-scores of each of the 38 classes of crop diseases. The chart reflects very strong performance, especially in classes like tomato (Healthy) and corn (Common rust).

Elbow curve (K-means clustering)

The elbow curve in Fig 6 indicates that for K=5, the intra-cluster variance is at a minimum for the K-means algorithm.

Confusion matrix for CNN model

The normalized confusion matrix (Fig 7) reflects the classification results of every class of plant diseases, showing a high performance for the classes such as tomato (Healthy) and corn (Common rust).

Silhouette score visualization for K-means

As shown in Fig 8, the mean silhouette score of 0.3647 still indicates moderate clustering quality, hence there is a need to improve cluster separation.

Residual plot for random forest model

The residual plot (Fig 9) shows that the random forest model does not have significant bias with respect to its predictions, further confirming its effectiveness.

2D PCA scatter plot (Cluster separation)

Fig 10 depicts the separation of the K-Means clusters in the reduced 2D feature space using PCA.

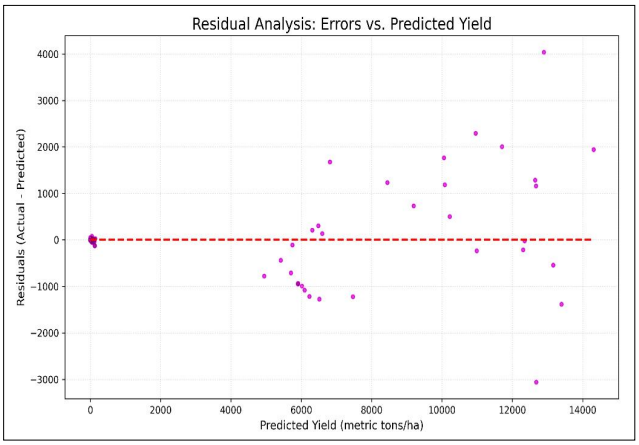


Fig 9: Residual plot (Errors vs. predicted crop yield).

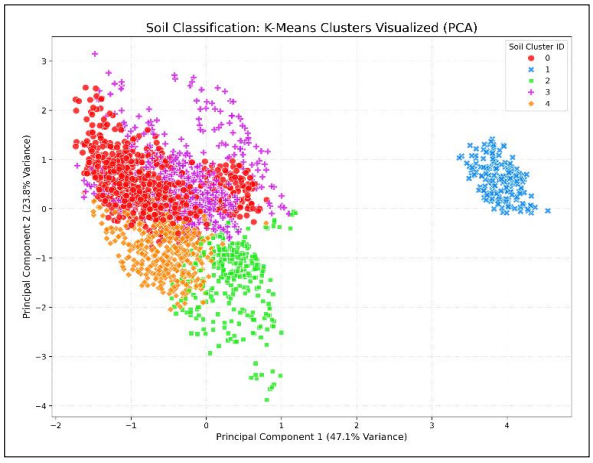


Fig 10: 2D PCA scatter plot for K-means clustering.

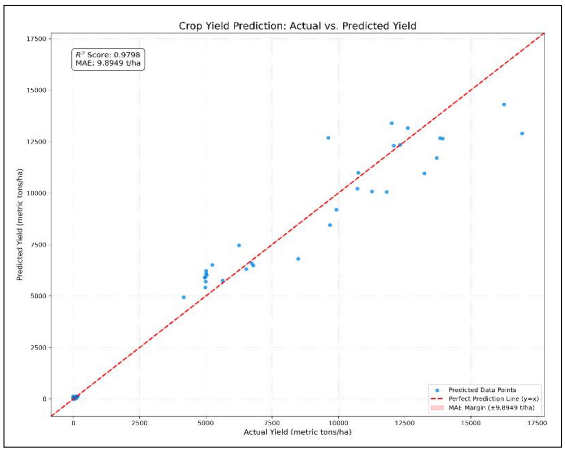


Fig 11: Predicted vs. actual crop yield.

Predicted vs. actual crop yield (Random forest)

The scatter plot in Fig 11 compares the predicted crop yields against actual crop yields, showing the high degree of accuracy the Random Forest model provides in predictions of agricultural outcomes.

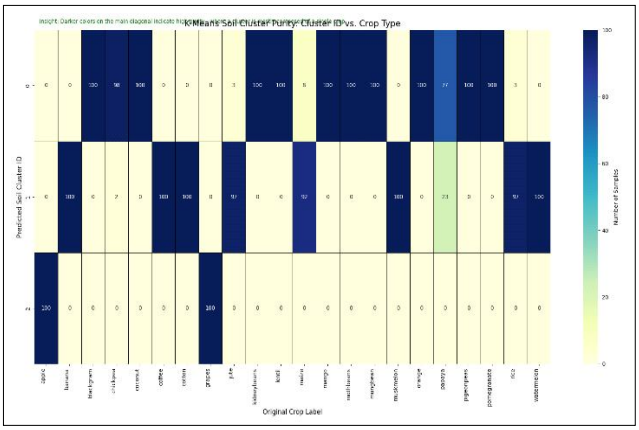


Fig 12: K-means cluster association heatmap (K=5).

K-means cluster association heatmap

Fig 12 shows the relationship between soil property and cluster assignment, which enables one to obtain insight into the chemical composition of different types of soil.

Result interpretation

The experimental outcomes validate the efficacy of the proposed multi-modal framework. The Random Forest model provides highly reliable yield forecasts (R^2= 0.9797), while the CNN demonstrates robust, generalized field diagnostics across 38 classes. Although K-Means achieved a moderate Silhouette Score, it establishes a highly acceptable baseline for processing high-variance soil chemistry. Together, these integrated models form a highly capable foundation for precision agriculture.

Future work

Some future enhancements could be the incorporation of live weather, satellite and IoT sensor data to enable real-time monitoring and yield prediction with greater accuracy. Further development of easy-to-use mobile or web tools, along with model customization for specific crops and regions, will enable farmers to make quicker, environment-friendly, data-driven decisions for sustainable and smart farming.

CONCLUSION

This paper proposes a machine learning-based framework to improve agricultural productivity by making decisions based on data evidence. It integrates the use of random forest regression for crop yield prediction, K-Means clustering for soil classification and a CNN for disease detection. The models produced very impressive results: R^2 = 0.9797, MAE = 9.47 kg/ha, CNN accuracy = 87.69%, Silhouette Score = 0.3647, proving that this framework can be used for yield forecasting, soil analysis and monitoring of crop health with good accuracy to assist in efficient and sustainable farming.

ACKNOWLEDGEMENT

First and foremost, the authors would like to thank the Head of the Department, Department of Computational Intelligence, SRM Institute of Science and Technology, for continuous support by motivating and guiding them.

Funding

The authors declare that this research was self-funded. No financial support or grant was received from SRM Institute of Science and Technology or any other funding agency in the public, commercial, or not-for-profit sectors.

Disclaimers

The views and conclusions expressed in this article are solely those of the authors and do not necessarily represent the views of their affiliated institutions. The authors are responsible for the accuracy and completeness of the information provided, but do not accept any liability for any direct or indirect losses resulting from the use of this content.

Informed consent

Not applicable (Study utilized public datasets and did not involve human or animal subjects).

Conflict of interest

The authors declare that there are no conflicts of interest regarding the publication of this article. No funding or sponsorship influenced the design of the study, data collection, analysis, decision to publish, or preparation of the manuscript.

REFERENCES

Abadi, M., Paul, B., Jianmin, C., Zhifeng, C. andy, D. and Jeffrey, D. (2016). TensorFlow: A System for Large-Scale Machine Learning. Proc. 12th USENIX Symp. Operating Systems Design and Implementation (OSDI'16). pp: 265-283.

AIRCC/AgriResearch (India). (2024). Random Forest Application for Crop Yield Prediction. Agri Research Journal.

Bhat, R., Joshi, R. and Arya, V. (2022). Crop production prediction models in indian agriculture: Possibilities and challenges. *Indian Journal of Ecology.* **49(3):** 1005-1010.

Cherukupalli, M.R. (2023). Crop yield prediction in Indian agriculture using machine learning. *AIP Conference Proceedings*. https://doi.org/10.1063/5.0169784.

Chollet, F. (2015). Keras: Deep Learning Framework. GitHub Repository. Available: https://github.com/keras-team/keras.

Dey, S., Goel, R.K., Chaurasiya, A.K. and Bania, R.K. (2026). Generation and utilization of an augmented cashew leaf dataset for disease analysis using transfer learning. *Indian Journal of Agricultural Research*. **60(3):** 428-435. doi: 10.18805/IJARe.A-6479.

Gharpure, A.A., Jain, N. and Narawade, V.E. (2026). Vit-CNN fusion for robust mango quality evaluation based on classification across multiple public datasets. *Indian Journal of Agricultural Research.* **60(3):** 436-442. doi: 10.18805/IJARe.A-6493.

Hughes, D.P. and Salathé, M. (2015). An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing. *arXiv preprint arXiv*. pp 1511.08060.

Kalmani, H.V., Dharwadkar, V.N. and Thapa, V. (2025). Crop yield prediction using deep learning algorithm based on CNN-LSTM with attention layer and skip connection. *Indian Journal of Agricultural Research.* **59(8):** 1303-1311. doi: 10.18805/IJARe.A-6300.

Kamilaris, A. and Prenafeta-Boldú, F.X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*. **147:** 70-90.

Keerthana, M., Meghana, K.J.M., Pravallika, S. and Kavitha, M. (2021). An Ensemble Algorithm for Crop Yield Prediction. Proc. 3rd Int. Conf. Intelligent Communication Technologies and Virtual Mobile Networks (ICICV). 963-970. doi: 10.1109/ICICV50876.2021.9388479.

Nikhil, U.V., Pandiyan, A.M., Raja, S.P. and Stamenkovic, Z. (2024). Machine learning-based crop yield prediction in South India: Performance analysis of various models. *Computers.* **13(6):** 137. https://doi.org/10.3390/computers13060137.

Patil, R., More, A. and Gatade, A.T. (2026). Analytical evaluation of CNN and capsulenet architectures for grape leaf disease prediction. *Indian Journal of Agricultural Research.* **60(3):** 443-449. doi: 10.18805/IJARe.A-6505.