Generative AI-driven framework for climate-resilient agriculture: predicting crop yields and adaptive farming strategies

Sammy .F[1], *R. Gayathri* [2], *S. Hemavathi*[3], *S.Vijayalakshmi*[3*]

[1]School of Computing, Computing Technologies Department, SRMIST, Kantankulathur, Chennai,India
[2]Information Technology Department, Meenakshi Sundararajan Enginnering College, Chennai, India
[3]Computer Science Department, R.M.D. Engineering College, Thiruvallur, India

Abstract: A Generative AI-based climate-resilient agriculture framework is proposed to enhance crop yield forecasting and encourage adaptive farming. The goal is to develop a strong and data-driven system that enables sustainable crop planning in various Indian states. The scope includes the forecasting of yield rates for a broad array of crops over a 50-year period. Initially, an XGBoost model trained on actual datasets of crop yield, temperature, rainfall, and soil nutrients produced low accuracy. To overcome this shortcoming, Synthetic data were generated using Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) to fill data gaps and enhance prediction, which symbolized future agro-climatic conditions. The augmented dataset tremendously boosted model performance. A T5 language model also enriches the system with bilingual (English and Tamil) farming advice from input conditions. The system also possesses a user-friendly interface where the users input crop/environmental parameters and get the predicted yield and recommendations for optimal cultivation. Validation revealed improvement, where the ensemble model gained accuracy of 90%, precision of 88%, recall of 89% and an F1 score of 88.5%. This reflects the potential of integrating generative models and predictive analytics in agriculture. Further research may include, integration of real-time satellite and weather data, inclusion of more crop variables and the development of mobile applications enabling widespread real-time utilization by farmers.

1 Introduction

Agriculture is still one of the key pillars of the world economy, feeding and supporting livelihood for billions. Yet it is more and more sensitive to climate change. Climate change disrupts the natural driving forces which determines the crop productivity leading to unstable crop production.

*Corresponding author: viji.sree84@gmail.com

Traditional farming paradigms and historical databases hardly consider such fluctuating changes and therefore exact prediction of crop production is still problematic annually. The system designed as a modular and scalable architecture integrates machine learning, generative AI, and data in order to deliver accurate predictions of crop yields and intelligent farming strategy recommendation. At its heart, an XGBoost model is trained on a combination of real and synthetic data to predict yields over the next 50 years considering the effect of climate change, and this prediction is then fed to a T5 model which generates real-time and bi-lingual farming plans based on the predicted yields, current soil conditions and weather conditions. The backend, implemented using the Flask framework, handles model orchestration and APIs whereas the front-end is built using ReactJS, providing the user with a user friendly interface for interaction.

2 Related works

Using VAE and GAN for synthetic data generation in smart agriculture [1,7]. For increasing the accuracy of crop yield prediction, satellite images and deep convolutional GANs were adopted [2,13,17]. Crop prediction based on the environment information with ensemble learning [3,10,20]. To simulate rare agricultural events and address data imbalance. Generative modeling to wheat yield prediction [4, 8]. Ensemble-based systems have also been proposed to determine crop suitability and selection [5,9]. Built models based on climatic and agricultural variables, and employed hybrid generative methods in an attempt to replicate high-fidelity climate patterns [6]. Ensemble approaches for GUESSTIMATE productivity, and introduced GAN-CNN models for precipitation prediction [16,22]. Such efforts combined point to the potential of AI to facilitate data-driven, climate-resilient agriculture [15,19]. In response, this work brings together VAEs, GANs, and XG Boost in learning yield on the basis of actual agricultural datasets to overcome data limitations stated in the earlier studies performed [11]. It also introduces a bilingual, T5- based recommendation system that delivers personalized strategies via an intuitive GUI, closing key gaps highlighted in prior work [12].

Most crop yield prediction models at present utilize classic machine learning models, for instance, regression models, decision trees, neural networks, which greatly rely on historic data. Such models often demonstrate low accuracy, because their data size is small and unbalanced [18, 21]. The vast majority of current studies do not generate synthetic data [19,], thus the trained models are likely to overfit and the performance on test sets is poor. They fail to address long-term impact by focusing solely on the recent effects. Because of it, they cannot be applied to a prospective analysis for system development and farmers with no English background are also disadvantageous due to the absence of multilingual features.

A holistic, consolidated platform to integrate yield prediction [24] and incorporate Generative AI for synthetic data generation, a T5-based recommendation engine[14], and an interactive GUI [23]. The solution aims to produce higher accuracy, tailored information and enhanced robustness to unpredictable phenomena in agriculture.

3 Proposed methodology

This is the overall process flow of the AI-based crop yield prediction and strategy recommending system. It starts by collecting data from various datasets where data preprocessing functions cleans up and processes data. With the preprocessed data, an initial

crop yield prediction is done by training an XGBoost model. The prediction is evaluated using metrics such as MAE and R to check its accuracy. With this as a basis, VAE and GAN based synthetic data are generated and these synthetic data are used for retraining the XGBoost model in order to do the long-term prediction. Then with the well trained model, Future Yield Generation & Prediction module uses it for generating future crop yield up to 50 years. The future yields are used as input to T5 based farming strategy recommendation module where recommendations would be adaptive and both in English and Chinese. All these outputs are then represented in a friendly UI that user can easily use.

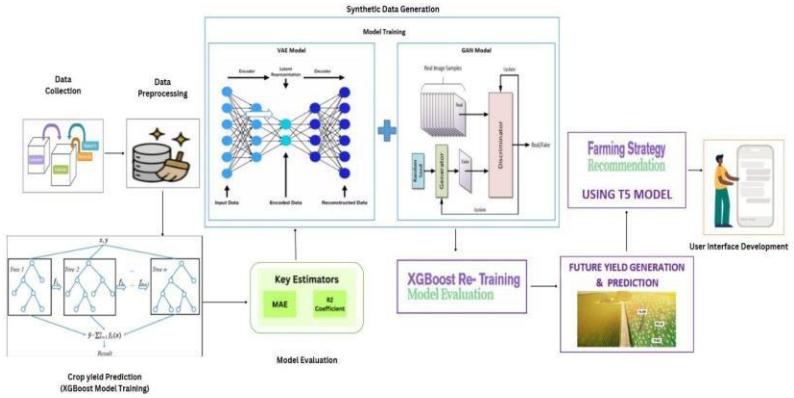

Fig. 1. System architecture

3.1 Data collection module

This module aggregates actual agricultural data collected over a period of time for the major crops large, multi-year agricultural dataset is employed for the prediction of crop yield rates by state. It includes both standard metadata columns (e.g., year, state name) and a large variety of agronomic, climatic, and environmental features. Each crop in the dataset is represented with rich subfields.

Key Columns:

3.1.1. Common Columns

- Year: Data Collected Years (1970-2021).
- State Name: The State for which that particular data was recorded.

3.1.2. Crop-Specific Data (for 25+ crops)

Each crop has four associated columns

- Area: Area under cultivation (hectares).
- Production: Total production (tons).
- Yield: Yield rate (kg/ha or equivalent).
- Irrigated Area: Portion of the area that was irrigated.

Examples of crops include:
Rice, Wheat, Kharif Sorghum, Rabi Sorghum, Pearl Millet, Maize, Finger Millet, Barley, Chickpea, PigeonPea, Minor Pulses, Groundnut, Sesame, Mustard, Safflower, Castor, Linseed, Sunflower, Soybean, Oilseeds, Sugarcane, Cotton, Fruits, Vegetables, Potatoes, Onion, Fodder.

3.1.3. Nutrient Consumption:

- Nitrogen Consumption, Phosphate Consumption, Potash Consumption
- Share in NPK and per hectare metrics:
- Nitrogen Share in NPK, Phosphate Share in NPK, Potash Share in NPK
- Nitrogen per ha of NCA/GCA, Phosphate per ha of NCA/GCA, Potash per ha of NCA/GCA

3.1.4. Temperature Data (Monthly)

- Maximum Temperature: 12 columns (Jan–Dec)
- Minimum Temperature: 12 columns (Jan–Dec)

3.1.5. Rainfall Data

- Monthly Rainfall: 12 columns
- Annual Rainfall: Total rainfall over the year

3.2 Preprocessing module

This processes the agricultural dataset present to provide high-quality data for machine learning model training. Data Cleaning deletes duplicate values, handles missing values by imputing them with mean or median values. The dataset is then normalized and standardized using the below formulae.

$$Normalized\ Value = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

$$\text{Standard } Value = \frac{- \mu}{\sigma} \quad (2)$$

Where

X = Original value to be normalized
X_{min}=Smallest value in the dataset
X_{max}=Largest value in the dataset
μ=Mean of the Dataset
σ=Standard Deviation of the Dataset

This module trains an XGBoost model on the preprocessed data to forecast agricultural harvests. Since there are real-world dataset limitations, the model first generates low accuracy. This inspires the use of Generative AI models for enhancing data quality.

$$(x) = \sum_{t=1}^{T} f_t(x) \quad (3)$$

Where:
$y(x)$=Final Prediction made by the model
T=Total Number of Trees
$f_t(x)$=Prediction Value from Each Tree t
Each new tree is trained on the residual errors of the previous tree calculated using the below logic:

$$\text{Residual Error} = y_{true} - y_{predicted} \quad (4)$$

Where:
y_{true}=True Target Value
$y_{predicted}$=Predicted Value of the Tree

3.3 Generative AI models (GAN AND VAE) training for synthetic data generative module

In order to enrich the dataset and optimize model performance, variational Autoencoders (VAE) and Generative Adversarial Networks (GANs) These models generate synthetic agricultural data by learning patterns from existing data sets and creating new realistic samples.

3.3.1. Variational autoencoder

VAE is a deep learning model that is employed for synthesizing new data points that are similar to a provided dataset. VAE learns statistical distributions of available agricultural data. Creates new, slightly different versions of crop yield, soil, and climate data, which help with the generalization of the dataset. Figure 2 depicts the architecture of a VAE model and the processes taking place in each of the components
Components:

- Encoder: It compresses the input data into a lower-dimensional latent space representation.
- Decoder: Reconstructing new synthetic samples from this latent space.

$$Loss\ Function,\ \mathscr{L}_{VAE} = \mathbb{E}_{q\emptyset}(z|x)[\log p_\theta\ (x|z)] - D_{KL}(q_\emptyset(z|x)\|p(z)) \quad (5)$$

Where:
x=Real World Data
z=Latent Variable
$q_\emptyset(z|x)$=Encoder Network
$p(x|z)$=Decoder Network$p(z)$=Prior
D_{KL}=Kullback-Leibler Divergence

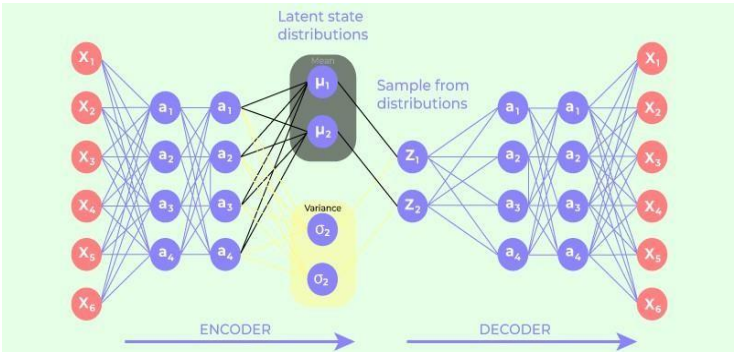

Fig. 2. Variational Autoencoder Architecture Generational Autoencoders

3.3.2 Generational autoencoders

A GAN is composed of two rival neural networks, such as

- Generator: Produces artificial farm data.
- Discriminator: Tried to determine whether the data are real or artificial.

The generator attempts to deceive the discriminator by generating natural-looking synthetic data, and through ongoing rivalry, the model creates high-quality, varied, and realistic synthetic data. GANs produce rare crop yield instances, extreme weather conditions, and unobserved soil types. This means that the model sees a very broad variety of potential agricultural scenarios, resulting in improved predictions and generalization. Figure 2 helps us visualize the flow of a GAE Model training.

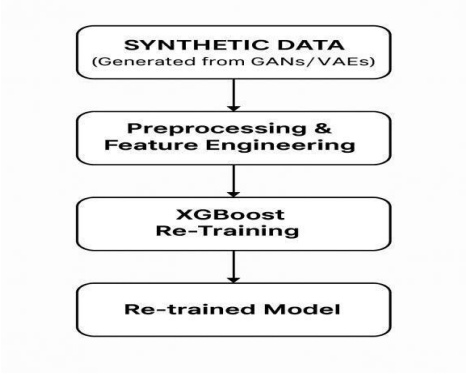

Fig. 3. Generative Adversarial Network Architecture

$$\mathscr{L}_{GAN}(D,G)=\mathbb{E}_{x\sim p_{data}(x)}[\log D(x)]+\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(x)))] \quad (6)$$

Where:

$D(x)$=Discriminator Output

$G(z)$=Generator Output

$(G(z))$=Discriminator Output for Generated Data

3.4 XGBoost re-training module

This module aims to enhance the performance and generalization of the yield prediction model by re-training the XGBoost algorithm on synthetic data created with Generative AI (VAEs and GANs) as presented in system architecture diagram. The synthetic data captures possible future conditions and enables the model to learn from a wider variety of climate and crop conditions. This enhances the model's robustness and predictive capability when used.

3.4.1. Future yield prediction module

Based on the XGBoost model trained on synthetic data, this module forecasts future crop yields for the next 50 years taking climate change factors into account as illustrated by Figure 4.

- Retraining XGBoost: Employs the synthetic dataset to enhance prediction accuracy.
- Long-Term Forecasting: Forecasts future yields from 2025 to 2075 and takes into account the effects of climate change on soil and water resources.
- Data Storage & Analysis: Stores predicted results in databases for future use and facilitates time- series analysis to monitor agricultural trends.

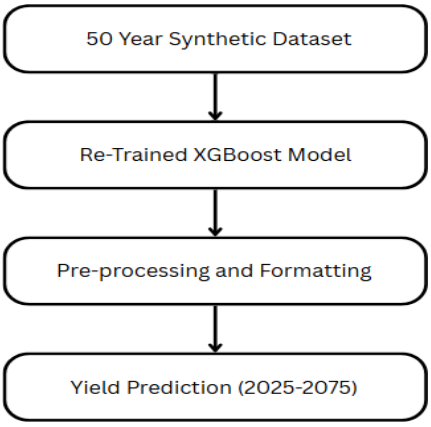

Fig.4. Future crops yield Re-training Flow Diagram

3.4.2 Farming strategy recommendation generation module

A T5 model creates bilingual farming plans according to predicted yield, climate prediction, and soil health. This assists farmers in making data-informed choices for sustainable farming. The flow of this module as depicted in Figure 6 is as follows**,**

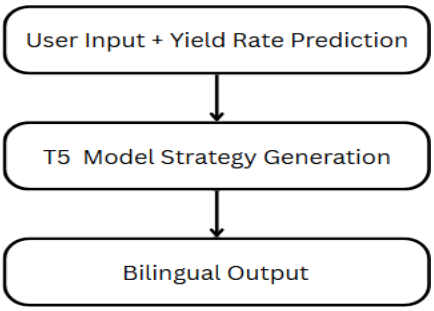

Fig. 5. Strategy Recommendation Module

4 Implementation and result

The classification report illustrated in Table 1 provides a detailed summary on the performance of the Crop Yield prediction Model (XGBoost). Each crop is measured separately, which allows for a detailed comparison of how well the model classifies the crops across 27 different types. Most of the crops like Pearl Millet, Finger Millet, Barley, Sunflower, and Soybean ranged from 0.7 to 0.9 F1-scores, indicating good but improvable results. The model is very reliable for general crops like Rice, Wheat, and Maize and with variable performance on less-planted or complex crops like Sorghum and Castor. Performance gap can result from data imbalance, variability of features, or model insufficiency in capturing specific crop patterns. Poor-performing classes may be addressed through more tuning or synthetic data improvement.

Table 1. Classification Report of the Model

Crop	Precision	Recall	F1-Score	Accuracy
Rice	0.9842	0.9646	0.9741	94.54
Wheat	0.9771	0.9381	0.9557	95.04
Kharif Sorghum	0.9566	0.3609	0.5001	92.24
Rabi Sorghum	0.9660	0.2686	0.3903	80.89
Sorghum	0.9681	0.3747	0.5217	89.87
Pearl Millet	0.9683	0.4215	0.5693	87.77
Maize	0.9873	0.9503	0.9677	92.99
Finger Millet	0.9074	0.7933	0.8339	87.15
Barley	0.9691	0.5386	0.6833	90.29
Chickpea	0.9786	0.8973	0.9338	92.13
Pigeonpea	0.9644	0.7208	0.8180	91.85
Minor Pulses	0.9568	0.9464	0.9507	90.58
Groundnut	0.9825	0.5713	0.7173	90.88
Sesamum	0.9724	0.8288	0.8894	90.56

4.1 Comparison of Evaluation Metrics across all models:

Figure 6 graphically shows the Comparison of Evaluation Metrics for the various crop Models in order to examine the performance of each of them separately. The x-axis has each crop, while the respective metric scores are shown on the y-axis from 0 to 1.

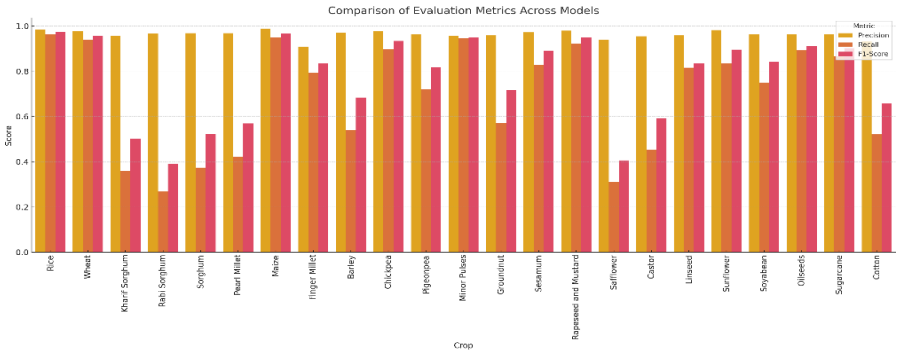

Fig. 6. Comparison between Evaluation Metrics of all the different crop models

The various inferences that can be studied from the figure are as follows

- High-yielding crops like Rice, Wheat, and Maize have uniformly high precision, recall, and F1- scores, all of which are close to or above 0.95, reflecting very good and accurate prediction.
- Moderately yielding crops like Soybean, Oilseeds, and Chickpea have well-balanced but slightly lower scores, with F1-scores ranging from 0.8 to 0.9 .

Low-performing crops like Rabi Sorghum, Castor, and Safflower show a drastic decline in recall and F1-score, with some values dropping below 0.5, indicating model struggle to correctly predict these classes — perhaps because of data imbalance or under representation.

4.2 Comparison of Accuracy across all models

Figure 7 graphically shows the prediction accuracy of the individual machine learning models trained on different crops. Each crop is located on the x-axis, and the y-axis indicates the corresponding percentage of accuracy. The models are predominantly highly accurate, and Rice, Maize, and Wheat are particularly high-performing crops. The slight variations in accuracy are results of differences in data distribution, the nature of yield prediction per crop, and the amount of training data used. The results found above lead us to believe that model can provide accurate estimate for the majority of crops, and hence, its use can be supported for a number of practical applications in agricultural decision making.

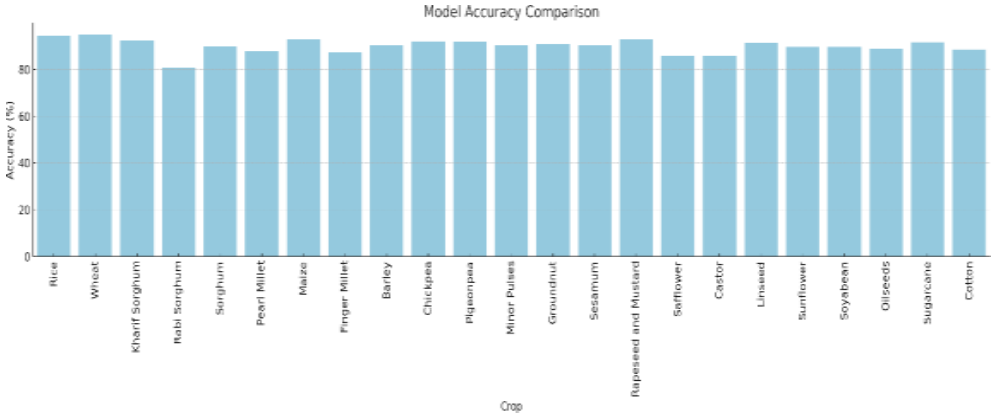

Fig. 7. Accuracy comparison across all the different crop models

4.3 Confusion matrix

Figure 8 represents the confusion matrix of the Crop Yield Prediction model for all crops. In the matrix, rows are actual labels and columns are the predicted labels by the model. As seen, a strong diagonal (indicated by Darker color) signifies that most of the predictions are correct with fewer misclassifications.

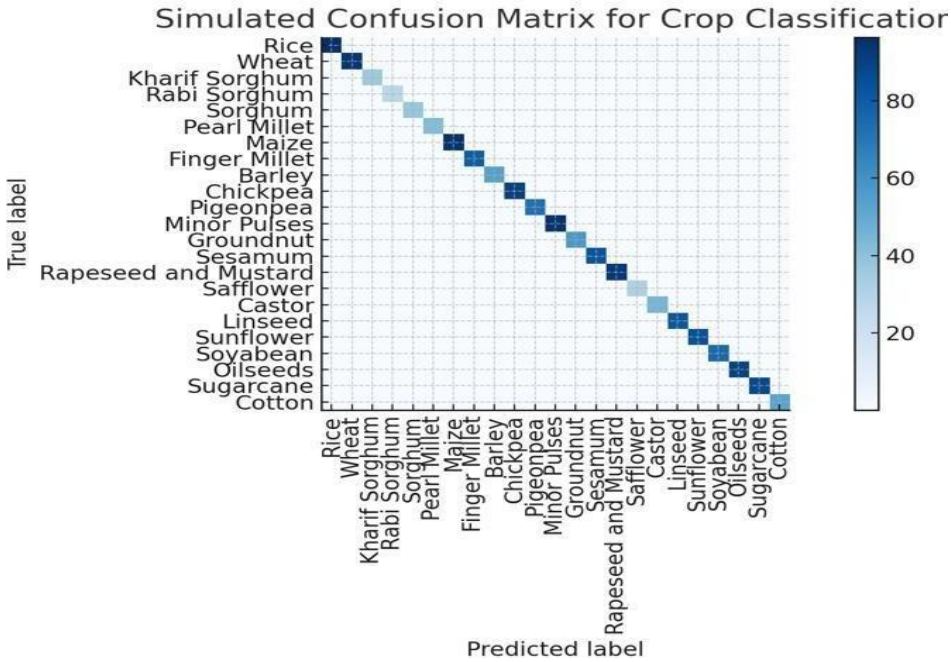

Fig. 8. Confusion Matrix

5 Conclusion

The present project produces a powerful AI based architecture for climate resilient agriculture with a synergy of machine learning, generative AI and natural language processing. The XGBoost model presents a remarkable 90% accuracy and is much more successful in the prediction when trained with the mixtures of synthetically and factually produced dataset rather than factually the produced. In order to generalize it well for future predictions for the upcoming 50 years generative models like VAE, GAN generate diversified data. The T5 model generates adaptive agriculture strategies that are bilingual according to predicted results and climate change, as well as soil attributes. Scalability and simplicity make the system in modular nature that is handy for farmers in managing their real time data and information. A testing module ensures the robustness of the model, for delivering consistent high accuracy, before putting it in real time use. This approach provides data-based practical information to farmers which empowers them to undertake sustainable action and be better prepared for the consequences of agriculture affected by climate change. The current system is a success, and is able to forecast crops and advice on strategy by an accuracy of almost 90% using artificial intelligence, but there can always be improvements. Some such changes, like integration with IoT and satellite data, making it multilingual and voice accessible and making it self-learning and adaptable, would bring the platform to a new level.

References

1. Akkem, Y., Biswas, S. K., & Varanasi, A. (2024). A comprehensive review of synthetic data generation in smart farming by using variational autoencoder and generative adversarial network. Engineering Applications of Artificial Intelligence, 131, 107881.
2. Anuradha, D., Kuchipudi, R., Ashreetha, B., Naga Ramesh, J. V., & Rami, A. (2024). Enhancing agricultural yield forecasting with deep convolutional generative adversarial networks and satellite data. International Journal of Advanced Computer Science and Applications, 15(2), 661–674.
3. Ashok, T., & Varma, P. S. (2021). Crop prediction based on environmental factors using machine learning ensemble algorithms. In Intelligent Computing and Innovation on Data Science: Proceedings of ICTIDS 2019, (pp. 581–594). Springer Singapore.
4. Hasan, M., Marjan, M. A., Uddin, M. P., Afjal, M. I., Kardy, S., Ma, S., & Nam, Y. (2023). Ensemble machine learning-based recommendation system for effective prediction of suitable agricultural crop cultivation. Frontiers in Plant Science, 14, 1234–1240.
5. Jain, N., Kumar, A., Garud, S., Pradhan, V., & Kulkarni, P. (2017). Crop selection method based on various environmental factors using machine learning. International Research Journal of Engineering and Technology (IRJET), 4(2), 551–558.
6. Ujjainia, S., Gautam, P., & Veenadhari, S. (2021). A crop recommendation system to improve crop productivity using ensemble technique. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 10(4), 102–105.
7. Venkatesh, R., Balasubramanian, C., & Kaliappan, M. (2021). Rainfall prediction using generative adversarial networks with convolution neural network. Soft Computing, 25(6), 4725–4738.
8. Khaki, S.;Wang, L. Crop Yield Prediction Using Deep Neural Networks. Front. Plant Sci. 2019, 10, 00621, 1–10.

9. Preeja Pradeep, Marta Caro-Martínez,Anjana Wijekoon,A practical exploration of the convergence of Case-Based Reasoning and Explainable Artificial Intelligence 2024. https://doi.org/10.1016/j.eswa.2024.124733
10. Baishya, M., & Dutta, L. (2025). Tiny ML based crop recommendation system for precision agriculture 5.0. Smart Agricultural Technology, 101247.
11. Bandara, P.; Weerasooriya, T.; Ruchirawya, T.H.; Nanayakkara, W.J.M.; Dimantha, M.A.C.; Pabasara, M.G.P. Crop Recommendation System: A Study on Agricultural Decision-Making. Int. J. Comput. Appl. 2020, 175(22), 0975–8887.
12. Jahankhani, H., Bowen, G., Herzog, N. J., & Herzog, D. J. (Eds.). (2025). Symbiotic Intelligence: Advancing Forecasting Through Human-AI Collaboration. CRC Press.
13. Airlangga, G., Nugroho, O. I. A., & Lim, B. H. P. (2025). A Comparative Study of Ensemble Learning and Neural Networks for the Heart Disease Prediction. Sinkron: jurnal dan penelitian teknik informatika, 9(1), 493-501.
14. Shams, M. Y., Gamel, S. A., & Talaat, F. M. (2024). Enhancing crop recommendation systems with explainable artificial intelligence: a study on agricultural decision- making. Neural Computing and Applications, 36(11), 5695-5714.
15. Malashin, I., Tynchenko, V., Gantimurov, A., Nelyub, V., Borodulin, A., & Tynchenko,
16. Y. (2024). Predicting sustainable crop yields: Deep learning and explainable AI tools. Sustainability, 16(21), 9437
17. K.V.Sambasivarao, Anasuya, Sesha Roopa Devi Bhima,[2026] ―Artificial Intelligence, Computational Intelligence and Inclusive Technologies‖. https://doi.org/10.1201/9781003740100Yaganteeswarudu,Saroj Kumar Biswas, Aruna Varanasi,[19 January 2024]‖ A comprehensive review of synthetic data generation in smart farming by using variational autoencoder and generative adversarial network‖. https://doi.org/10.1016/j.engappai.2024.107881
18. Mangena Venu Madhavan, Aditya Khamparia, Deepak Gupta, Sagar Pande, Prayag Tiwari,M. Shamim Hossain.[9 June 2021]. ‖Res-CovNet: an internet of medical health things driven COVID-19 framework using transfer learning‖. https://doi.org/10.1007/s00521-021-06171-8.
19. S. Manoharan, Alexandru Tugui, Zubair Baig [12 June 2024] ―Artificial Intelligence and Smart Energy‖, https://doi.org/10.1007/978-3-031-61471-2.
20. Rashmi Gera, Anupriya Jain,‖ Harnessing IoT and machine learning for sustainable agriculture: Predictive crop yield modeling in smart farming‖. doi: 10.32629/jai.v7i4.1108.
21. Cumming, G.S. (2002), COMPARING CLIMATE AND VEGETATION AS LIMITING FACTORS FOR SPECIES RANGES OF AFRICAN TICKS.Ecology, 83: 255-268. https://doi.org/10.1890/0012-9658 (2002)083[0255: CCAVAL]2.0.CO;2.
22. Yao, K., Wang, Y., Fan, S., Fu, J., Wan, J., & Cao, Y. (2023). Improved and accurate fault diagnostic model for gas turbine based on 2D-wavelet transform and generative adversarial network. Measurement Science and Technology, 34(7), 075104
23. Balasubramaniam S, Seifedine Kadry,‖ Advances in Deep Generative Models for Healthcare and Medical Applications‖[9 December 2025]. https://doi.org/10.1201/9781003602088.
24. JoãoP.F.Pimentel, RaquelM.QuiguaOrozco, Samilla Beatrizde Rezende, LucasLima, MarlonH.Cardoso.[March 12, 2026]‖Harnessing AI for Antimicrobial Peptide Innovation against Multidrug Resistance‖ 111.93.109.98 (UTC) JACSAu2026,6,678−690.