Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 8 (2026) 3291

Journal of the Nigerian Society of Physical Sciences

CropGAN: A conditional GAN framework for synthetic tabular data augmentation in crop recommendation systems

Dekera Kenneth Kwaghtyo, Christopher Ifeanyi Eke*, Timothy Moses

Department of Computer Science, Faculty of Computing, Federal University of Lafia, P.M.B. 146, Lafia, Nigeria

Abstract

Crop recommendation systems play a crucial role in precision agriculture by enabling informed decisions about which crops to cultivate in a specific location. However, their performance is often hindered by the inherent scarcity, imbalance, and regional bias of tabular agricultural datasets. These limitations reduce the reliability and generalizability of crop recommendation models, especially in data-scarce regions. Existing synthetic data generation methods, such as the Synthetic Minority Oversampling Technique (SMOTE) and variational autoencoders (VAEs), struggle to handle high-dimensional, structured agricultural data with categorical variables. Moreover, existing generative adversarial networks (GANs) are primarily image-focused. This study proposes a re-engineered GAN, termed the Crop Recommendation GAN (CropGAN), to generate tabular, multi-crop recommendation data using a class-conditioning mechanism. The framework was designed to handle complex, non-linear datasets with both numerical and categorical variables, addressing dataset size, imbalance, and limited diversity in multi-crop datasets. CropGAN was trained on a dataset of 5,000 samples with 10 crop classes and evaluated against SMOTE and VAE using statistical data-quality and classification-performance metrics. In terms of data-quality assessment, SMOTE best preserved the original data distributions, while CropGAN introduced greater diversity in the synthetic datasets. In terms of classification performance, models trained on CropGAN-generated samples consistently achieved the highest performance, with the support vector machine (SVM) yielding the best accuracy of 99.4%. This result suggests that adversarial learning can be adapted for tabular agricultural datasets to improve the performance of crop recommendation systems in data-scarce regions.

DOI: 10.46481/jnsps.2026.3291

Keywords: Generative adversarial network, Synthetic tabular data, Data imbalance, Precision agriculture, Crop recommendation

Article History:
Received: 3 February 2026
Received in revised form: 24 April 2026
Accepted for publication: 14 May 2026
Available online: 24 June 2026

Communicated by: P. Thakur

1. Introduction

Crop recommendation models analyse soil and climatic attributes to support data-driven decisions on which crops to plant at specific locations [1, 2]. In particular, they use variables such as soil pH, nitrogen (N), phosphorus (P), potassium (K), temperature, and rainfall to suggest crops likely to thrive in given locations and environmental conditions [3–5]. However, the performance of crop recommendation models solely performance depends on dataset size, balance, and diversity[2]. Conversely, crop recommendation datasets are usually scarce and limited in size, highly imbalanced, and less diverse, thus limiting the generalizability and reliability of such models [6].

In the Yandev district of Gboko local government, Benue State, Nigeria, the majority of the people are small-scale farmers who cultivate less than two hectares of farmland [7].

*Corresponding author Tel. No.: +234 706 809 0013
Email address: eke.christopher@science.fulafia.edu.ng (Christopher Ifeanyi Eke)

As such, these dataset limitations are particularly pronounced, leading to minority classes like pepper and yams often containing fewer than 50 instances, compared to rice with over 500 instances. This results in a dataset imbalance ratio exceeding 1:10. This imbalance often introduces bias in model training to the detriment of the minority classes, which affects model reliability [2, 7].

Several approaches, ranging from oversampling techniques to generative models, have been developed to address the challenges of data scarcity and class imbalance. Despite these efforts, these techniques have accompanying limitations. For instance, the Synthetic Minority Over-Sampling Technique (SMOTE) addresses data imbalance by linearly interpolating between the minority class instances. Though effective for handling imbalance, SMOTE is limited in capturing complex, non-linear soil and climatic features, leading to duplication of data samples [8, 9]. Variational autoencoders (VAEs) generate synthetic samples via a latent compression process. The imposed latent compression or regularisation often levels out the class labels, limiting its effectiveness in multi-class crop recommendation datasets [9, 10].These limitations motivated the need to explore generative frameworks with the inherent ability to handle complex dataset features while preserving their class labels.

The advent of Generative Adversarial Networks (GANs), which uses an adversarial training process with the ability to handle complex, non-linear and linear dataset features, has brought a paradigm shift in data augmentation. GANs have exhibited strong performance in agricultural image-based tasks like crop and weed segmentation, pest and disease detection [11–14]. However, most implementations of GAN in agriculture focus on image-based datasets. These models may not be directly adopted to augment crop recommendation datasets, which are text-based comprising categorical and numerical instances. The lack of a tailored GAN framework for crop recommendation datasets represents a significant research gap, especially in data-scarce regions where data augmentation can improve crop recommendation models.

To address this challenge, this study re-engineered the traditional GAN architecture, now referred to as Crop-recommendation GAN (CropGAN), for the synthetic generation of multi-crop recommendation datasets. CropGAN modifies the traditional GAN framework by integrating a class-conditioning mechanism for multi-class processing, capable of handling complex, non-linear features. The purpose is to address dataset scarcity, imbalance and lack of diversity of multi-crop recommendation datasets. This synthetic generation of realistic and balanced dataset samples by the CropGAN framework significantly enhances the performance of crop recommendation models, useful in data-scarce regions.

The specific contributions of this study remain:

i. Adaptation of the traditional GAN architecture to generate crop recommendation datasets with both numerical and categorical features.
ii. Integration of class-aware mechanism for handling class imbalance in multi-crop recommendation datasets.
iii. A detailed evaluation of CropGAN against alternative methods (SMOTE and VAE) using both statistical data-quality and model performance metrics.
iv. Empirical validation of CropGAN using both ML and DL models for crop recommendation tasks in data-scarce regions.

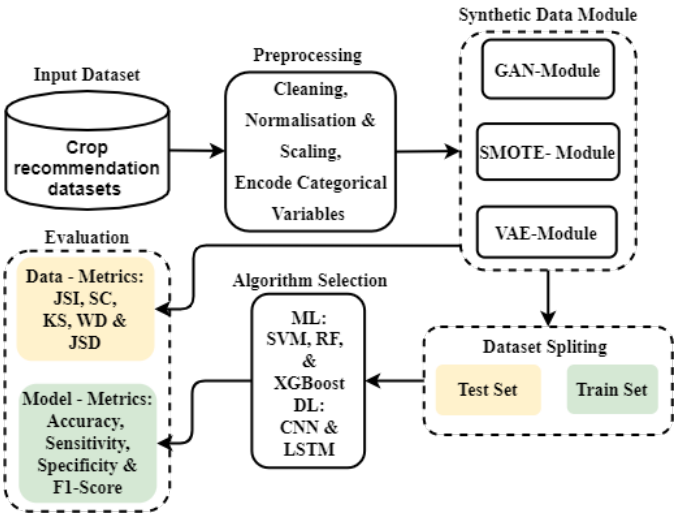

Figure 1: Model architecture.

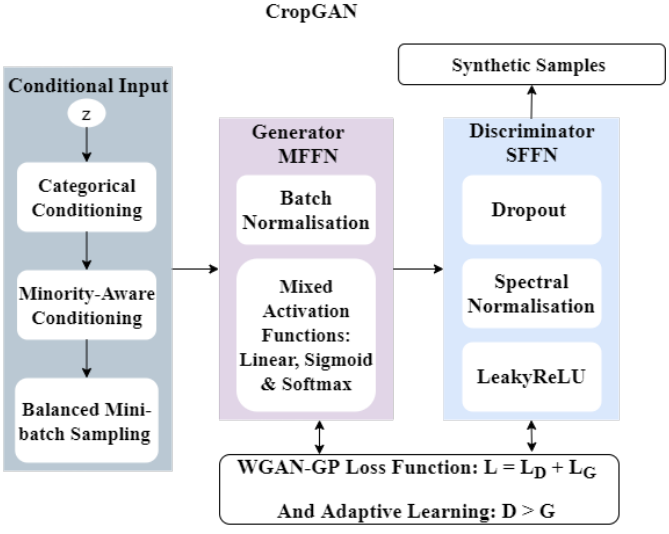

Figure 2: The CropGAN architecture.

The remainder of the work is organised into four sections: Section 2 focuses on the empirical review relating to the study. Section 3 discusses the methodology, including the experimental settings and procedures. Furthermore, section 4 presents and discusses the experimental results of CropGAN. Lastly, section 5 provides a recap of the study and concludes with specific recommendations for future work.

Table 1: Sample of the original dataset utilised in the study.

	N	P	K	temp	hum	ph	rainf	soil	label
0	40	27	45	21.66	94.79	5.89	112.43	sandy loam	guinea_corn
1	120	7	47	24.25	83.04	6.65	54.77	loamy clay	pepper
2	102	11	47	27.99	92.78	6.50	27.15	sandy loam	tomatoes
3	5	25	6	30.72	94.01	6.01	106.81	loamy	orange
4	36	43	21	28.36	84.86	7.14	52.93	loamy	yam
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4995	120	20	45	25.67	88.70	6.11	54.23	loamy clay	pepper
4996	36	55	20	27.01	84.34	6.64	55.30	loamy	yam
4997	37	56	25	22.06	19.60	5.77	126.73	clay	beans
4998	68	57	43	26.09	80.38	5.71	182.90	clay	rice
4999	88	17	52	29.90	90.75	6.65	25.38	sandy loam	tomatoes
5000 rows × 9 columns									

Table 2: Sample of the CropGAN generated dataset.

	N	P	K	temp	hum	ph	rainf	soil	label
0	85.00	58.00	41.00	21.77	80.32	7.04	226.66	0.0	6.0
1	3.00	9.00	45.00	23.89	89.62	6.54	104.62	3.0	2.0
2	42.00	67.00	77.00	18.99	15.94	6.11	78.71	1.0	7.0
3	32.00	13.00	42.00	23.50	92.98	5.79	106.62	3.0	2.0
4	106.00	21.00	52.00	28.90	94.79	6.29	23.04	3.0	8.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
5995	0.35	0.01	0.14	0.12	0.14	0.05	0.29	0.0	9.0
5996	0.08	0.20	0.13	0.10	0.09	0.10	0.12	0.0	9.0
5997	0.01	0.00	0.12	0.16	0.10	0.04	0.21	0.0	9.0
5998	0.03	0.29	0.03	0.06	0.48	0.02	0.12	1.0	9.0
5999	0.17	0.01	0.03	0.26	0.47	0.11	0.43	0.0	9.0
10000 rows × 9 columns									

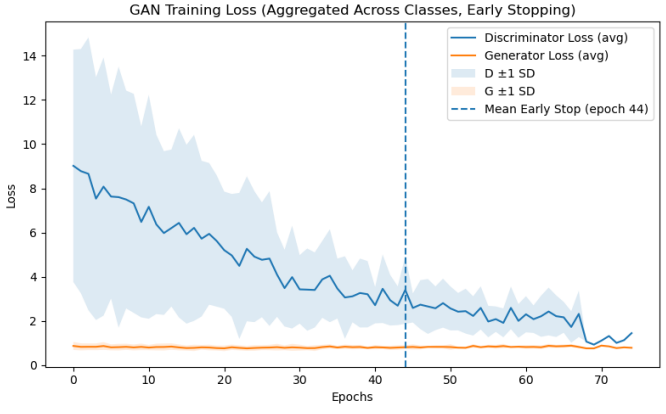

Figure 3: CropGAN training loss plot.

2. Empirical studies

The empirical review of existing studies on synthetic-data-generating models was thoroughly examined in this section. The existing synthetic approaches considered in the review include SMOTE, VAE and GAN models, often used to enhance precision agricultural systems. These approaches are thematically reviewed as presented in the subsequent subsections.

2.1. SMOTE-enhanced models in precision agriculture (PA)

The problem of dataset imbalance has been a persistent challenge across several fields, including agriculture. Innovative solutions to address this challenge led to the advent of the Synthetic Minority Over-sampling Technique (SMOTE) approach. SMOTE interpolates the minority classes to produce artificial data samples, thereby addressing class imbalance and increasing sample size. Therefore, this section reviews existing works on SMOTE and its variants as it establishes the base for this study. In this regard, Sarkar, Zhou, Scaboo, Zhou, Aloysius and Lim [15] developed a framework for addressing imbalance in image datasets by combining the Synthetic Minority Oversampling Technique and Edited Nearest Neighbours (SMOTE-ENN). SMOTE-ENN was applied to an image dataset comprising 1,266 samples to evaluate the breeding condition of soybean lodging. The Extreme Gradient Boosting (XGBoost), Random Forest (RF), K-Nearest Neighbour (KNN), and Artificial Neural Network (ANN) were trained on the dataset. Results indicate that the model achieved 96% accuracy, demonstrating the effectiveness of the hybrid oversampling method. Cuenca-Romero, Apolo-Apolo, Rodríguez Vázquez, Egea and Pérez-Ruiz [16]

Table 3: Models hyperparameter configurations.

Model	Parameters and values
Random forest	Number of trees = 50; Maximum depth = 2, and Random_state = 42
Support vector machine (SVM)	Kernel = RBF; Regularisation parameter(C) = 0.03; and Random_state = 42
XGBoost	Number of estimators = 5; Maximum depth = 2; Learning rate = 0.5; Subsample = 0.2; Column subsample = 0.2; Evaluation metric = mlogloss; and Random state = 42
Convolutional neural network (CNN)	Number of convolution layers =2; Hidden units = {64, 128}; Kernel size = 3; Dropout = 0.6; Dense layer = 64; Optimizer = Adam(lr = 0.001); Batch size = 32; Maximum epochs = 500(early stopping)
Long Short-Term Memory (LSTM)	LSTM layers = {64, 32}; Dropout = 0.7; Dense layer = 64 units; Activation = ReLU, Dropout = 0.7), Optimizer = Adam; Batch size = 32; and Maximum epochs = 500 (early stopping)
CropGAN (Generator/Discriminator)	Generator: Dense(128, activation = ReLU, BatchNormalization(), Dense(64, activation= ReLU), BatchNormalization(), Dense(input_dim, activation = linear); Discriminator: Dense(64, activation = ReLU, Dropout1(0.3), Dense(32, activation = ReLU), Dropout(0.3), Dense(1, activation = sigmoid), Loss = Wasserstein loss with gradient penalty(WGAN-GP), optimizer = Adam, lr = 0.0001, β_1 = 0.5 and β_2 = 0.9.)
Variational Autoencoder (VAE)	Encoder: dense layers = 2 (64, 128 units, ReLU), latent_dim =10, β = 0.001; Decoder: dense layers = 2 (64, 128 units, ReLU, sigmoid), optimiser = Adam, lr = 0.001, Batch size =32, epochs=500
SMOTE (for data augmentation)	sampling_strategy = target_samples_per_class (1000), k_neighbors = 1, random_state = 42

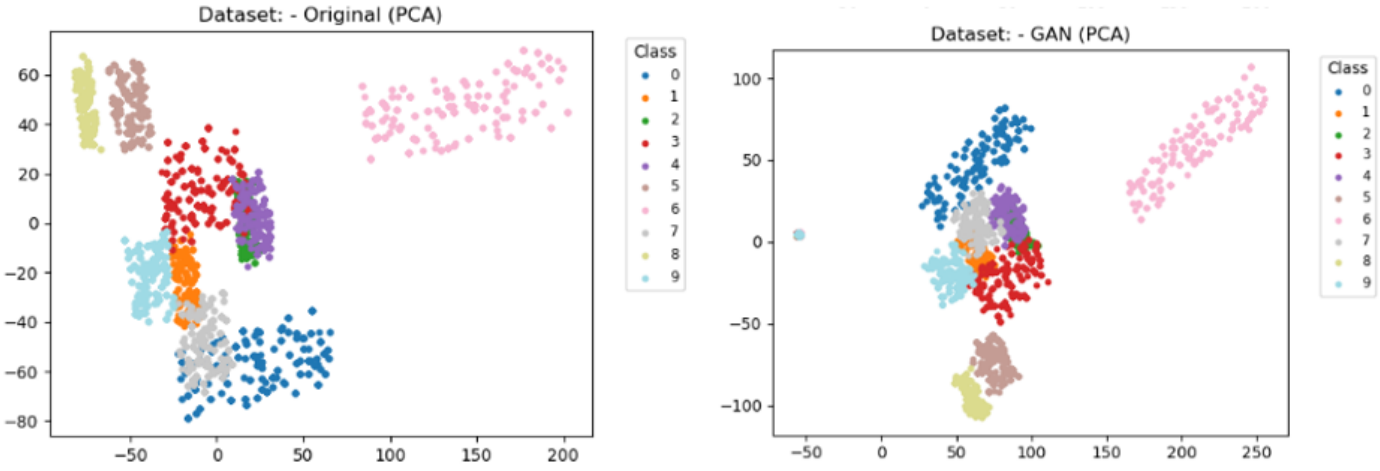

Figure 4: The PCA plots for the original and CropGAN datasets.

Table 4: Statistical metrics analysis.

Dataset	JSI	SRC	KSS	WD	JSD
Original	1.0000	1.0000	0.0000	0.0000	0.0000
SMOTE	1.0000	1.0000	0.0290	0.0480	0.0950
VAE	1.0000	1.0000	0.0890	0.1440	0.1450
CropGAN	1.0000	1.0000	0.0940	0.1670	0.1610
Hybrid-SMOTE	1.0000	1.0000	0.1350	0.1740	0.3450
Hybrid-VAE	1.0000	1.0000	0.1450	0.1860	0.3470
Hybrid-GAN	1.0000	1.0000	0.1220	0.1540	0.3170

employed SMOTE to balance hyperspectral data to improve the detection of yellow and brown rust in wheat farms. Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forests (RF), and Gaussian Naïve Bayes (GNB) were

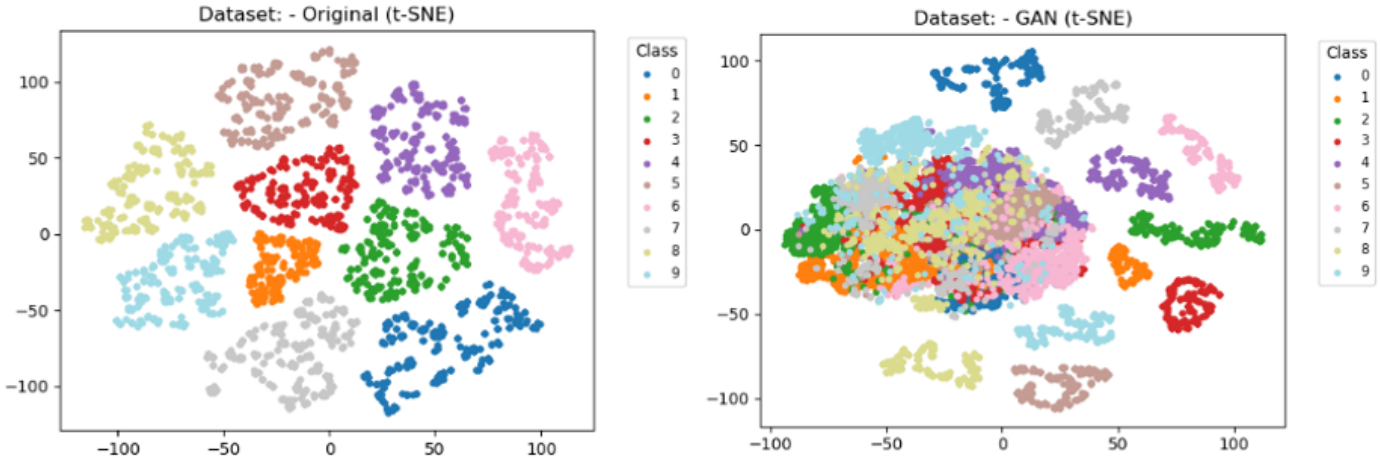

Figure 5: The t-SNE plots for the original and CropGAN datasets.

Table 5: Performance summary of models across dataset types.

Dataset	Model	Accuracy	Precision	Recall	F1-score
Original	SVM	0.985	0.9866	0.985	0.9843
	XGBoost	0.966	0.9714	0.966	0.9656
	LSTM	0.930	0.9480	0.930	0.9258
	RF	0.940	0.9009	0.940	0.9163
	CNN	0.895	0.9265	0.895	0.8843
CropGAN	SVM	0.994	0.9942	0.994	0.9939
	RF	0.986	0.9871	0.986	0.9860
	XGBoost	0.979	0.9818	0.979	0.9788
	LSTM	0.972	0.9743	0.972	0.9696
	CNN	0.926	0.9529	0.926	0.9188
Hybrid-CropGan	SVM	0.985	0.9866	0.985	0.9843
	XGBoost	0.966	0.9714	0.966	0.9656
	LSTM	0.955	0.9630	0.955	0.9445
	RF	0.940	0.9009	0.940	0.9163
	CNN	0.781	0.7444	0.781	0.7213

trained on the SMOTE-enhanced dataset. Evaluation results indicate that for yellow rust detection, RF outperformed with 70% accuracy, while SVM outperformed with 63% accuracy for detecting brown rust. Senapaty, Ray and Padhy [17] used SMOTE to address class imbalance and trained Logistic Regression, Decision Tree, K-Nearest Neighbour, Support Vector Machine, Random Forest, Gradient Boosting, Bagging Tree, XGBoost, AdaBoost, CatBoost, Histogram-Based Gradient Boosting, Stochastic Gradient Descent Classifier (SGDC), and Multinomial Naïve Bayes algorithms. Experimental results indicate that SGDC achieved an excellent result of 1.00 accuracy, with sensitivity and specificity values of 0.91 and 0.54, respectively. Similarly, Iorzua, Kwaghtyo, Hule and Ibrahim [2] employed SMOTE to mitigate class imbalance and used the balanced dataset to train Random Forest, Decision Tree, Naïve Bayes, Logistic Regression, and Extreme Gradient Boost (XGBoost) algorithms for recommending suitable crops to plant. Experimental results show that XGBoost outperformed the rest of the algorithms with 99.65% accuracy. Rahat, Ghosh, Dara and Kant [13] applied SMOTE to address class imbalance using the pixel values extracted from 4,000 image datasets to train the CNN model. Evaluation results demonstrate that the SMOTE-enhanced framework outperformed other models, having 99% accuracy, while precision and recall yielded 0.95 and 0.98, respectively. Sapkal and Kadam [18] applied SMOTE to mitigate class imbalance to enhance the performance and robustness of a crop recommendation system. Experimental results indicate that the XGBoost model yielded the best performance of 93.82% accuracy.

2.2. *VAE-enhanced models in precision agriculture*

Attempts to address the limitations of the SMOTE method led to the introduction of the Variational Autoencoders (VAEs) method for addressing data scarcity and imbalance. This section, therefore, reviews existing studies on VAE-enhanced models in precision agriculture as it forms the background for this study. Consequently, Iatrou, Karydas, Tseni and Mourelatos [19] used a Variational Autoencoder (VAE) to address dataset

Table 6: Ablation results for benchmarking CropGAN performances.

Dataset	Model	Accuracy	Precision	Recall	F1-score
SMOTE-only	LSTM	0.998	0.9980	0.998	0.9980
	SVM	0.987	0.9893	0.987	0.9872
	XGBoost	0.961	0.9683	0.961	0.9604
	RF	0.927	0.9671	0.927	0.9188
	CNN	0.677	0.6634	0.677	0.6219
Hybrid-SMOTE	SVM	0.985	0.9866	0.985	0.9843
	XGBoost	0.966	0.9714	0.966	0.9656
	LSTM	0.953	0.9658	0.953	0.9454
	RF	0.940	0.9009	0.940	0.9163
	CNN	0.589	0.5731	0.589	0.5055
VAE-only	SVM	0.991	0.9922	0.991	0.9911
	CNN	0.989	0.9896	0.989	0.9890
	LSTM	0.981	0.9822	0.981	0.9811
	XGBoost	0.975	0.9784	0.975	0.9747
	RF	0.937	0.9693	0.937	0.9327
Hybrid-VAE	LSTM	0.967	0.9737	0.967	0.9631
	SVM	0.985	0.9866	0.985	0.9843
	XGBoost	0.966	0.9714	0.966	0.9656
	RF	0.940	0.9009	0.940	0.9163
	CNN	0.802	0.8757	0.802	0.7596

imbalance and scarcity to enhance the estimation of nitrogen requirements in a rice farm to improve yield. The algorithms Catboost, XGBoost and LightGBM were trained on the augmented dataset. Performance evaluation indicates that the VAE-augmented dataset enhanced the performance of the model, yielding a Mean Absolute Error (MAE) of 0.6 tn/ha, outperforming existing approaches. Cao, Ma and Zhang [20] applied VAE to augment remote sensing data for corn yield analysis. The augmented dataset was used to train the developed VAE-based Multiple Instance Regression (VAEMIR) model. Experimental analysis indicates that the proposed VAEMIR model achieved an R^2 value of 0.74, outperforming existing approaches. Razavi, Nejadhashemi, Majidi, Razavi, Kpodo, Eeswaran, Ciampitti and Vara Prasad [21] used VAE to address the dataset imbalance and scarcity challenge to improve the performance of crop yield models. The Random Forest, XGBoost, CatBoost, and LightGBM algorithms were trained on the VAE-augmented dataset. Experimental results demonstrate that the CatBoost model achieved RMSE of 0.330, nRMSE of 0.093, MAE of 0.17, and MASE of 0.485, on the original dataset. While on the VAE synthetic data, the CatBoost model achieved 93.6%, surpassing other models. Isinkaye, Olusanya and Akinyelu [6] combined VAE and Vision Transformers (ViTs) to preprocess and augment image datasets to enhance plant disease detection. Performance evaluation indicates that the hybrid model classified plant diseases, achieving a high performance of 93.2% accuracy. Liu and Song [22] combined Temporal Convolutional Network (TCN) and VAE approaches to tackle dataset limitations, including scarcity, imbalance and missing values for precise crop yield analysis. Evaluation results demonstrate that the hybrid TCN–VAE framework outperformed traditional ML and DL models, yielding an RMSE of 0.245, MAE score of 0.192, and a 0.935 R^2 value. This result surpasses the LSTM model, which achieved an RMSE of 0.298 and 0.908 R^2, including the GRU model that yielded 0.287 RMSE, 0.915 R^2 value, among others.

2.3. GAN-enhanced models in precision agriculture

The inability of SMOTE and VAE approaches in handling complex, non-linear and linear relationships in multi-crop recommendation datasets leaves an important research gap. Generative Adversarial Networks (GANs) emerged as a powerful tool for addressing data-related challenges in various domains, including agriculture. In this section, GAN-enhanced models in precision agriculture are reviewed to establish the need for their variants tailored for augmenting multi-crop recommendation datasets. As a result, Karam, Awad, Abou Jawdah, Ezzeddine and Fardoun [23] employed GAN to augment image datasets to enhance the performance of their pest detection model. Evaluation analysis shows that GAN-augmented data increased recall performance from 54.4% to 93.2% at 0.50 IoU. Bird, Barnes, Manso, Ekárt and Faria [24] used Conditional Generative Adversarial Network (CGAN), a GAN variant, to create synthetic images to address scarcity and imbalance challenges. A transfer learning model (VGG16) was trained on the augmented dataset to enhance fruit quality analysis. Experiments using 2,690 images yielded 83.77% accuracy, while the GAN-augmented dataset increased the classification performance to 88.75% accuracy. Shumilo, Okhrimenko, Kussul, Drozd and Shkalikov [25] employed GAN to tackle class imbalance in satellite imagery for crop and soil analysis. Performance experiment shows that the model's performance increased by 1.5% accuracy with Intersection Over Union (IoU) of 2.1%. Most importantly, the minority crop classification

results significantly increased by 11.2% accuracy. Kwong, Kon, Rusik, Shabudin, Rahman, Kulaveerasingam and Appleton [26] used GAN to handle data scarcity and imbalance to improve an oil palm detection model. Evaluation results indicate that the model achieved 95.8% and 97.2%, respectively, on precision and recall metrics. Fawakherji, Suriani, Nardi and Bloisi [12] developed a GAN variant tagged "Deep Convolutional Generative Adversarial Network (DCGAN)" to create synthetic images, thereby addressing the limitations of size and imbalance. Experimental evaluations demonstrate that the synthetic method achieved 94% performance accuracy. Wang, Xia, Xia, Wang and Gu [14] introduced a GAN variant named "Frequency-Domain and Wavelet Image Augmentation Network with a Dual Discriminator Structure (FHWD)" to address data scarcity in crop disease detection. Experimental results show that the FHWD-augmented dataset improved the performance of disease classification by 7.25% when applied to transfer learning models such as VGG16, GoogleNet, and ResNet18.

3. Methodology

The method employed to adapt or re-engineer the traditional GAN architecture tailored for the synthetic generation of multi-crop recommendation datasets is described in five (5) main stages. The core stages include: (i) Dataset acquisition; (ii) Data cleaning/preprocessing; (iii) CropGAN module; (iv) Classifier module; and (v) Evaluation module. Figure 1 represents the entire methodology.

3.1. Dataset acquisition

The dataset used in this study was originally collected by Kwaghtyo, Eke, Abah, Moses, Agushaka and Fatokun [7]. The dataset initially comprised 2200 samples, but the dataset was expanded to 5000 samples to effectively experiment with the present study. The initially collected 2200 samples are retained to ensure comparability with previous findings, while the additional 2800 new samples were collected. The new samples were collected using a random sampling approach across farm fields in the Yandev district of Gboko Local Government Area (LGA), Benue State. This sampling was carried out across two farming seasons (2024 and 2025) to gain seasonal variability. The soil samples were collected between 0 cm and 30 cm depth using soil augers. These samples were subjected to laboratory procedures to obtain the soil nutrient composition and physiochemical properties. In terms of climatic conditions, temperature ranges from 25.00°C to 33.50°C, while rainfall ranges from 900 to 1,200 mm. The dataset is made of nine attributes, including Nitrogen (N), Phosphorus (P), Potassium (K), Soil pH, Humidity, Temperature, Rainfall, Soil Texture, and Crop Type. The crops considered include maize, rice, pepper, soybean, beans, orange, guinea corn, cassava, tomatoes, and yams, which are mostly cultivated in the area. These agronomic and environmental features together form essential input parameters for training the developed CropGAN synthetic data generation method to address the limitations of SMOTE, VAE and the traditional GAN in handling multi-crop recommendation datasets. Table 1 represents the dataset as viewed in a Jupyter notebook environment.

Table 1 shows the structural and feature-level composition of the original dataset, highlighting the mixture of both the numerical and categorical attributes. This represents/illustrates the nature of agricultural tabular datasets used in crop recommendation tasks.

To facilitate direct comparison with the original dataset in Table 1, Table 2 presents a representative sample from the synthetic dataset generated by the CropGAN framework. This allows for an expeditious assessment of structural consistency and feature coherence between the original and synthetic datasets.

As demonstrated in Table 2, the generated samples preserve the distributional structure of the original dataset, exhibiting realistic values consistent with the original dataset. Meanwhile, the representative samples generated using the SMOTE and VAE data augmentation techniques are included as Tables A.1 and A.2, respectively, in Appendix A.

3.2. Cleaning/pre-processing

Data cleaning and pre-processing are crucial steps to remove noise and irregularities, preparing the dataset for training a self-learning model. Therefore, the collected data underwent various cleaning and pre-processing procedures to improve its quality. Specifically, Exploratory Data Analysis (EDA) was performed to examine and detect anomalies such as merged text, missing values, and extraneous spaces. These data cleaning tasks were automated using Jupyter Notebook, a Python interactive environment, to ensure reproducibility. During EDA, missing values were addressed with median imputation to manage skewed data distributions. Feature scaling was carried out using Min-Max Normalisation to keep all features within a consistent range, aiding stability and convergence of algorithms. Categorical variables (crop_labels and soil_texture) were encoded with one-hot encoding. The EDA process revealed class imbalance, with some crop classes being significantly underrepresented. This issue, which motivates this study, was left to be tackled by the synthetic data generation methods used.

3.3. Synthetic data generation frameworks

The synthetic data generation module was developed to address the data limitations, including scarcity, imbalance and diversity. While GAN was re-engineered and tagged CropGAN, it was trained separately alongside SMOTE and VAE for the purpose of generating synthetic data points for comparison. The implementation of each of these architectures is discussed thus:

3.3.1. The CropGAN architecture

CropGAN is a specialised Generative Adversarial Network (GAN) developed to tackle the unique challenges of crop recommendation datasets. These datasets often include varied features, both numerical and categorical thus suffer from limited diversity and class imbalance. Such issues make it difficult for standard GANs, originally designed for image data, to function effectively. Typical GANs struggle to maintain

feature distributions or generate valid categorical data required in crop recommendations. To overcome these challenges, CropGAN incorporates three key modifications: (i) a class-aware mechanism that accurately synthesises mixed data types, (ii) a conditional embedding system with minority-aware sampling to address class imbalance, and (iii) replacing binary cross-entropy with Wasserstein Loss with Gradient Penalty (WGAN-GP) to improve training stability and reduce mode collapse. The architecture of CropGAN includes four main parts: the generator, discriminator, conditional input layer, and a specialised loss function, as illustrated in Figure 2.

i. The generator layer

The generator layer uses a Multi Feedforward Network (MFFN) to convert the latent noise vector and conditional inputs into realistic data samples. To improve training stability and gradient flow, each hidden layer incorporates batch normalisation and Leaky ReLU activations. The output layer handles mixed data types by using specific activation functions: sigmoid for binary variables and softmax for multi-class categorical variables. This strategy helps ensure that the generated samples preserve the statistical and structural features of the original data. The generator function is defined as follows:

$$\tilde{x} = G(z_c; \theta_G) = \sigma\left(W^{(L)} h^{(L-1)} + b^{(L)}\right), \tag{1}$$

$$h^{(L)} = \text{LeakyReLU}\left(\text{BN}\left(W^{(L)} h^{(L-1)} + b^{(L)}\right)\right), \tag{2}$$

where $W^{(L)}$, $b^{(L)}$ = respective weights and biases of layer L.
BN$(\cdot)$ = Batch Normalisation.
$\sigma(\cdot)$ = denotes the mixed activation functions (sigmoid and softmax).
$h^{(L)}$ = the hidden layer activation.

This design choice enables the generator to produce diverse but valid dataset samples while still preserving feature distributions.

ii. The discriminator layer

The discriminator layer is implemented as a Symmetric Feedforward Network (SFFN), mirroring the generator to provide a balanced adversarial learning. It receives both the original and the synthetically generated samples together with their corresponding conditional inputs and returns scalar scores that represent authentic data samples. To improve training stability and generalization, dropout regularisation and spectral normalisation are integrated in the discriminator layer. This helps in controlling overfitting. The discriminator function is given by:

$$D(x, c; \theta_D) = W^{(L^l)} \phi^{(L^l-1)} + b^{(L^l)}, \tag{3}$$

$$\phi^{(L)} = \text{LeakyReLU}\left(\text{Dropout}\left(\text{SN}\left(W^{(L)} \phi^{(L-1)} + b^{(L)}\right)\right)\right), \tag{4}$$

where $D(x, c; \theta_D)$ = The discriminator output.
$W^{(L)}$, $b^{(L^l)}$ = The weight matrix and bias vector of layer L.
SN = Spectral normalisation.
Dropout$(\cdot)$ = Regularisation.
$\phi^{(L)}$ = Activation of layer L.

This design configuration ensures a robust critic that is capable of determining the subtle variations between the original and the synthetically generated crop recommendation dataset.

iii. Conditional input layer

To ensure controlled dataset generation that tackles class imbalance, CropGAN integrates a conditional input mechanism. The class labels, $y \in \{1, \ldots, K\}$, are transformed into continuous embeddings: $e(y) = \text{EmbedLabel}(y) \in \mathbb{R}^{d_c}$.

These embeddings are then concatenated with the latent noise vector z, which is fed into both the generator and the discriminator.

To further ensure that class imbalance is addressed, minority-aware sampling is applied, where class labels are sampled from a reweighted distribution. These are represented mathematically as follows:

$$\tilde{x}_g = G_\psi(z, e(y)), \quad z \sim p_z(z), \quad y \sim \tilde{p}(y), \tag{5}$$

$$\tilde{p}(y) \propto \frac{1}{\left(\pi_y\right)^\gamma}, \quad \pi_y = \frac{N_y}{N}, \quad \gamma \in [0, 1], \tag{6}$$

where $\tilde{x}_g$ = Generated data conditioned on class embedding $e(y)$.
$p_z(z)$ = Prior distribution over the latent vector.
$\tilde{p}(y)$ = The reweighted class prior distribution.
$e(y)$ = Continuous embedding vector of class label y.
N_y = Number of training samples in class y.
N = Total number of training samples across all classes.

This mechanism ensures a targeted generation of minority class samples, thereby improving dataset balance to enhance model performance.

iv. The loss function

To enhance stability and avoid mode collapse during training, CropGAN adopts the Wasserstein loss with Gradient Penalty (WGAN-GP) instead of the conventional binary cross-entropy.

The CropGAN loss function is given by:

$$L_D = \mathbb{E}_{x \sim p_r(x|c)}[D(x|c)] - \mathbb{E}_{z,c}[D(G(z, c), c)] + \lambda \mathbb{E}_{\hat{x}}\left[(\|\nabla_{\hat{x}} D(\hat{x}, c)\|_2 - 1)^2\right], \tag{7}$$

with the gradient penalty terms:

$$\hat{x} = \epsilon x + (1 - \epsilon)\tilde{x}, \quad \epsilon \sim \mathcal{U}[0, 1], \tag{8}$$

where $\mathbb{E}_{x \sim p_r(x|c)}[D(x|c)]$ = expected critic (discriminator) score on the real data,
$\mathbb{E}_{z,c}[D(G(z, c), c)]$ = expected critic (discriminator) score on the generated data,
$\hat{x} = \epsilon x + (1 - \epsilon)\tilde{x}$, $\epsilon \sim \mathcal{U}[0, 1]$, where $\hat{x}$ denotes a real sample, $\tilde{x}$ denotes a generated sample and λ is the gradient penalty term enforcing the 1-Lipschitz constraint for stabilising training.

This is specifically important in crop recommendation datasets where instability can lead to severe effects on the

quality of the generated data samples.

v. Training procedure

CropGAN is trained using 500 epochs, Adam optimiser with a learning rate of 1×10^{-4}, $\beta_1 = 0.5$ and $\beta_2 = 0.9$. To prevent mode collapse, the discriminator is updated multiple times per generator step. Using these parameter settings, CropGAN generates 10,000 data samples from the original dataset comprising 5,000 instances, enhancing dataset diversity and preserving feature distributions.

vi. CropGAN contributions

In contrast to the conventional GAN, CropGAN is specifically designed to generate tabular agricultural datasets by: (i) supporting the generation of mixed categorical and numerical features using a class-aware mechanism, (ii) tackling class imbalance through conditional embedding and minority-aware sampling, (iii) improving training stability and sample quality using WGAN-GP, and (iv) capturing non-linear feature relationships critical for crop recommendation datasets. These design choices enable CropGAN to produce high-fidelity, diverse and structurally consistent synthetic datasets that enhance the robustness and performance of crop recommendation models.

3.3.2. The SMOTE architecture

To establish a baseline for evaluating CropGAN, the Synthetic Minority Oversampling Technique (SMOTE) was implemented as a comparative method. Unlike CropGAN, SMOTE performs simple feature-space interpolation. The implementation was done via Python's imblearn library in a Jupyter notebook environment. To ensure a balanced class distribution, each class was upsampled to 1,000 instances, which returned a total of 10,000 samples with all classes equally represented. The purpose of this fixed class sampling target was intended to ensure that the SMOTE synthetic dataset has the same number of instances as CropGAN and VAE to prevent a biased comparison due to unequal sample sizes. Each class label was mapped to its corresponding target sample count, thereby enforcing equal samples across all classes in the dataset. To minimise the generation of unrealistic samples, the number of nearest neighbours was set to 1, while the random_state of 42 was used to ensure reproducibility. The SMOTE algorithm was applied to the original training features ($X_{\text{train\_real}}$, $y_{\text{train\_real}}$), producing synthetic feature vectors ($X_{\text{train\_smote}}$, $y_{\text{train\_smote}}$). This was stored as a dataframe with the original feature names preserved for downstream modeling. The final augmented training set ($X_{\text{aug\_smote}}$, $y_{\text{aug\_smote}}$) was formed by concatenating the original and synthetic samples.

3.3.3. The VAE architecture

Another comparable synthetic data generation method to benchmark with CropGAN is the Variational Autoencoder (VAE). VAE was implemented using the TensorFlow/Keras library and trained to effectively capture class-specific latent feature distribution. Before training, all input features were normalised using MinMax scaling. The encoder generated a mean and log-variance vector for each input, facilitating stochastic sampling through reparameterisation, while the decoder reconstructed data samples from the latent representations. Optimisation was achieved using a composite loss function that combined mean squared reconstruction error and Kullback–Leibler (KL) divergence penalty, regulated by a beta (β) coefficient to balance reconstruction fidelity and regularisation of latent space. The VAE model was trained using 500 epochs, a 32-batch size, and the trained decoders synthesised additional data, producing a uniform target of 1,000 samples per class. The training result yielded an overall total of 10,000 synthetic data samples. This target sample size of 1,000 was intentional to maintain a balanced sample size with SMOTE and CropGAN-generated data to avoid biased comparison due to unequal sample size. The uniform sample size across SMOTE, VAE and CropGAN provided a fair ground for comparison across the implemented synthetic data generation methods.

3.4. The classification models

Several classifiers were used to assess the performance of the re-engineered GAN tagged CropGAN for generating synthetic multi-crop recommendation datasets. The classifiers were purposefully selected across ML and DL models to gain more insight into the comparative assessment of the effectiveness of CropGAN synthetic data. The ML algorithms employed include Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost). Particularly, RF was used due to its ability to handle non-linear features. SVM was employed to benefit from its ability to handle high-dimensional multi-class datasets. While XGBoost was leveraged for its regularisation and parallelisation mechanisms that handle noise and missing values in datasets, thereby improving the performance of crop recommendation models. Complementarily, the DL architectures Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) were employed. Although DL models are not originally designed for handling tabular datasets, they were included to gain more insight into their behaviour on the CropGAN synthetic dataset. The inclusion of both ML and DL frameworks enhances the comparative analysis of the effectiveness of the CropGAN-generated dataset on diverse models.

3.5. Evaluation procedure

To examine the behaviour of the synthetic data generation methods SMOTE, VAE, CropGAN, and their hybrid sets with the original dataset, both data-quality and model performance metrics were employed.

3.5.1. Dataset evaluation metrics

This study employs five (5) statistical metrics, such as Jaccard Similarity Index (JSI), Spearman's Rank Correlation (SRC), Kolmogorov–Smirnov Statistic (KSS), Wasserstein Distance (WD) and the Jensen–Shannon Divergence (JSD) to assess the distributional relationship of the synthetic datasets to the original dataset. The JSI assess the degree of overlap in the categorical feature distributions of the real and synthetic datasets. This technique ensures that the categorical variables (crop_labels and soil_texture) are preserved. SRC checks

whether the relationships among variables within the synthetic data type are preserved. Both KSS and WD check for distributional diversity and fidelity of features in the synthetic datasets and the original dataset. JSD assesses the symmetric divergence involving the real and synthetic data in terms of their probability distributions. Among these metrics, WD was prioritised for its effectiveness in capturing both diversity and fidelity in the synthetically generated datasets. While the JSI was chosen as a complementary metric for assessing categorical variables, which are very critical in crop recommendation tasks. Complementarily, visualisation metrics like Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbour Embedding (t-SNE) were used to respectively visualise the structural patterns and clustering behaviour of the synthetic and original datasets.

3.5.2. Model evaluation metrics

In terms of model performance, standard evaluation metrics such as Accuracy, Precision, Recall and F1-Score were used. Accuracy provides an overall measure of correct classifications, while Recall measures the ability of the models to correctly identify suitable crops to plant. Precision evaluates the ratio of correctly predicted crops among all the positive predictions, and F1-Score balances precision and recall. Precision was designated as the core performance evaluation metric in this study, due to its focus on minimising false positives in predictions. A wrong recommendation of crops may lead to resource wastage or low crop yield, jeopardising the core purpose of crop recommendation models. The choice of Precision is consistent with existing agricultural studies, where Precision has been prioritised in crop recommendation and crop disease detection tasks. In this study, Precision consistently produced superior performance relative to Accuracy and F1-Score, as reported in Tables 5 and 6. This thorough evaluation ensures a detailed assessment of both dataset quality and predictive performance, enabling objective analysis of whether synthetic data generated by CropGAN, SMOTE, and VAE effectively supports the development of reliable crop recommendation models.

3.6. Experimental settings and procedure

Experiments carried out on both the original and synthetic datasets generated using SMOTE, VAE and the CropGAN architecture followed standard processes. First, the dataset was partitioned into training (70%), validation (15%), and testing (15%) sets to maintain a consistent class distribution across all splits. The used conventional ML models, including RF, SVM and XGBoost, were all trained using the default optimisation strategies provided by Scikit-learn and XGBoost libraries, while model-specific hyperparameters were selected via empirical parameter tuning. The deep learning models (CNN and LSTM) were trained for a maximum of 500 epochs with early stopping based on validation loss to prevent overfitting. A batch size of 64 and Adam optimiser with a learning rate of 0.001 were used for all deep learning experiments. Cross-entropy loss was consistently applied for multi-classification across all models. Each model was trained and evaluated separately on the original and synthetic augmented datasets. Model performance was assessed using standard metrics such as accuracy, precision, recall, and F1-score. Additionally, dataset feature distribution similarity between the original and synthetic datasets was examined using statistical metrics like JSI, WD, KSS, JSD and SRC, while PCA and t-SNE were employed for visual inspection. All these experiments were conducted using a personal computer (PC) with Intel(R) Core(TM) i5, 12GB RAM, and a CPU speed of 2.60 GHz.

All implementations were developed using Python 3.9, while Scikit-learn and XGBoost libraries were used for ML models, while TensorFlow/Keras and PyTorch were used for implementing the DL and generative models, respectively. The data pre-processing was performed via Pandas and NumPy, while Matplotlib was employed for visual analyses. All these experiments were executed using the Jupyter Notebook environment. The best performing parameter tuning across all models is reported in Table 3.

As shown in Table 3, the parameters reflect a balance between model complexity and generalization ability. For the ML models, limited tree depth in both the RF and XGBoost were employed to prevent overfitting due to the dataset size. For the DL models, dropout regularisation and early stopping were used to enhance model generalization and representational capacity.

4. Results and discussion

4.1. Visual analysis

4.1.1. CropGAN training outcome

The CropGAN training process was stable, with the discriminator loss starting at around 9 and steadily decreasing to between 1 and 2; the generator loss relatively remained constant around 0.8 and 1.0. This indicates a balanced adversarial game without collapse. Early stopping was applied at approximately epoch 44, where the generator and the discriminator losses reached their minimum difference before convergence. Variability in loss was initially high but steadily reduced as training progressed, confirming convergence as shown in Figure 3.

4.1.2. The PCA plot

Principal Component Analysis (PCA) technique was employed to inspect the relationship between the original and the CropGAN-generated datasets, projecting the high-dimensional feature space into 2-dimensions. Figure 4 depicts the outcome of the PCA on both the Original and the CropGAN-synthetic data. However, the original dataset appeared more compact, with most classes having separate clusters. In contrast, the CropGAN-generated dataset preserves the overall distributional structure of the original dataset but exhibits increased dispersion, which results in some overlap between classes 2, 3, 4 and 5. However, class 9 maintains a clearly distinct cluster in both datasets. These observations show that CropGAN-generated data introduces variability (diversity) while maintaining the structure of the original dataset.

4.1.3. The t-SNE plot

The t-Distributed Stochastic Neighbour Embedding (t-SNE) was used to visualise the non-linear structure and local separability of the data features in both datasets. As shown in Figure 5, the original data forms a dense and well-separated cluster for most classes. The CropGAN-generated dataset exhibited a less dense cluster with an increased overlap between certain classes. This result aligns with the PCA analysis, confirming that the CropGAN-generated dataset preserves the overall structures but shows higher variability (diversity) with reduced local separability as compared to the original dataset.

4.2. Synthetic data quality performance outcome

The performance of the CropGAN-generated data was analysed and compared against SMOTE and VAE using five statistical metrics. These include the Jaccard Similarity Index (JSI), Spearman correlation, Kolmogorov-Smirnov Statistic (KS), Wasserstein distance, and Jensen-Shannon Divergence (JSD). Table 4 provides a summary of these results.

Table 4 depicts that SMOTE, VAE and CropGAN, including their hybrid datasets, all returned 1.0000 on both JSI and SRC metrics. This result signifies their ability to preserve the structural relationship of the original dataset. On the divergence-based metrics, SMOTE obtained 0.0290, 0.0480, and 0.0950 values for KSS, WD and JSD, respectively. This result suggests that the SMOTE-generated data preserved the original dataset distribution with minimal variations. VAE, however, produced slightly higher divergence with KSS, WD and JSD yielding 0.0890, 0.1440, and 0.1450, compared to SMOTE. This signifies that VAE introduces more divergence from the original dataset distribution than SMOTE. Comparatively, the CropGAN synthetic dataset demonstrated greater divergence with KSS (0.0940), WD (0.1670), and JSD (0.1610) values. This result confirms that CropGAN produces more diverse and realistic data samples in the synthetic dataset, required in modeling multi-crop recommendation systems [27]. In the hybrid datasets, the divergence slightly increased, indicating that the hybridisation process introduced some noise since the original dataset had zero divergence to contribute.

It is important to note that all the synthetic data generation methods preserve the distributions of the original dataset to varying degrees. SMOTE provides the closest distribution but creates less diverse data samples due to its interpolation policy, which may limit the generalizability of a model [28]. VAE and CropGAN produced results in terms of preserving the distributions of the original dataset and creating a more diverse dataset. The adversarial training policy of CropGAN can be likened to its ability to generate a more diverse dataset, capturing the complex, non-linear relationships which SMOTE (interpolation-based) and VAE (reconstruction-based) methods may struggle with.

4.3. Classifier performance outcome

Another dimension employed to assess the performance of the CropGAN method is the downstream performance of its generated dataset on model development. To achieve this, five algorithms including Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boost (XGBoost), Long-Short Term Memory (LSTM) and Convolutional Neural Network (CNN) were used. These models were trained using the Original, CropGAN and the Hybrid of the Original and CropGAN datasets, to gain more insight into their effect across models. Table 5 represents the performance of these various datasets across the used models.

As shown in Table 5, the CropGAN synthetic dataset slightly outperformed the original and the hybrid datasets across all models. This performance is evident with the SVM model, indicating that CropGAN synthetic dataset enhanced class separability, addressing class imbalance and data scarcity (size). The constant high performance of the SVM model across all datasets can be attributed to its margin-based optimisation, suitable for crop recommendation datasets, while benefiting from the smooth decision boundary introduced by the CropGAN samples. While the hybrid-CropGAN dataset also yielded a competitive performance result for several models, it was not uniform across all architectures. The ML models showed a consistently good performance on the hybridised samples; the CNN model, however, showed a drastic performance degradation. This result can be attributed to the inherent sensitivity of convolutional architectures to the feature ordering and spatial dependencies, with assumptions that are not aligned with crop recommendation datasets.

4.4. Ablation analysis

In this section, an ablation study was conducted to isolate and evaluate the CropGAN data generation framework on the overall classification performance. The ablation variables in the study are the alternative data generation methods. CropGAN was systematically replaced with alternative data generation methods to determine whether the performance improvement observed in Section 4.3 is only attributable to the CropGAN synthetic dataset or reproducible by other data generation methods. Consequently, four ablation conditions were considered: (i) SMOTE-only generated data, where all models were trained on the SMOTE-generated data; (ii) VAE-only dataset, under which models were trained exclusively on the VAE-only generated samples; (iii) Hybrid-SMOTE samples, where all models were trained using a combination of both the SMOTE-only and the original datasets; (iv) Hybrid-VAE samples, in which models were trained on a fused VAE-only and the original datasets. All the hybrid datasets were combined using simple row-wise concatenation, while shuffling was used to ensure uniform labels, with each having 10,000 samples as the CropGAN, SMOTE and VAE for a fair comparison. The outcome of the ablation study, summarised in Table 6, is a benchmarked result of the CropGAN performance shown earlier in Table 4.

As shown in Table 6, replacing CropGAN with SMOTE and the VAE datasets brought a noticeable performance variation across the used models. While SMOTE achieved high performance for a sequence-based model like LSTM, SMOTE's effectiveness degraded with CNN, suggesting limited robustness for addressing class imbalance in isolation. The VAE-only dataset on the other hand, demonstrates more stable per-

formance results across the classifiers, but its performance remains below that of the CropGAN framework. This indicates that the VAE's latent space reconstruction remains inadequate for handling complex nonlinear feature dependencies of crop recommendation datasets. In the SMOTE and VAE hybrid dataset variants, there are inconsistent improvements and, in some instances, their performance degraded. This indicates that the data fusion might have introduced some noise into the hybrid dataset rather than improving model generalization. These findings further confirm that the performance improvement reported in Section 4.3 is unique to the CropGAN framework, not generic to all synthetic data generation methods.

4.5. Discussion

This study examined the effectiveness of CropGAN, designed for the generation of multi-crop recommendation datasets. The focus of this discussion is on CropGAN training stability, the distributional properties of the synthetic datasets, their data quality and model performance impacts relative to VAE and SMOTE methods.

The performance of the CropGAN synthetic method during training, as visualised in Figure 3, highlights a stable adversarial learning process. PCA and t-SNE visualise the training outcome by showing the structural distribution/clustering. The slightly less compact clustering observed in Figure 4 and 5 may suggest that CropGAN introduced more diverse samples while maintaining the overall structure of the original dataset. This diversity can support model performance, generalizability and robustness.

Statistical analysis using the JSI and SRC metrics also yielded a strong correlation across all datasets. Divergence-based metrics such as the KSS, WD, and JSD further revealed distributional variations within each synthetic data sample. SMOTE yielded the lowest divergence relative to the original data. This can be attributed to its interpolation process, the possibility of replicating data samples rather than producing an entirely new sample. Comparatively, the VAE and CropGAN samples showed higher divergence compared to the original data distribution, with CropGAN being the highest. This indicates that the adversarial training process of CropGAN introduces more variability in the samples. The trade-off of distributional fidelity and diversity, where CropGAN preserved the overall structure with reduced clustering, aligns with the visualisations of the PCA and t-SNE plots.

Classification performance further validated the data-quality results, as all the used models, when trained on the CropGAN samples, consistently improved model performance compared to the original dataset. The superior performance of the SVM model across all data variants indicates that the introduced variation in the CropGAN sample improved class separability while reducing feature sparsity. However, the hybrid CropGAN variant produced a mixed performance result where the CNN model drastically degraded while other models maintained their performance. This finding suggests that the sensitivity of convolutional architectures to feature ordering is not well suited to crop recommendation datasets. Ablation analysis replacing CropGAN with alternative synthetic samples (VAE-only and SMOTE-only) achieved good performance for LSTM on the SMOTE-only samples, indicating robustness across both ML and DL architectures. The VAE-only sample yielded a more stable performance across models, but less than that of CropGAN. The hybrid variants of SMOTE and VAE exhibited inconsistent results and, in some instances, degraded in performance.

Overall, these findings suggest that the achieved performance by the CropGAN synthetic method is unique and not generic to all synthetic methods. The performance can be attributed to its adversarial training process that introduced diversity while still maintaining key distributional properties of the original data. By expanding the features, CropGAN supports model improvement for multi-crop recommendation datasets that comprise both numerical and categorical variables.

5. Conclusion

This study presents CropGAN, an adapted GAN architecture for the generation of synthetic datasets for modeling crop recommendation systems. The architecture was adapted to tackle dataset issues such as limited size, imbalance, and limited diversity. Unlike other GAN variants, CropGAN was specifically re-engineered to handle both categorical and numerical features by enforcing class-conditional input in the adversarial learning. Statistical and classifier analyses have shown that the CropGAN synthetic method produced realistic samples of the original dataset through its adversarial learning strategy. Although SMOTE was observed to retain a closer structural similarity to the original dataset, CropGAN consistently outperformed both SMOTE and VAE in terms of model performance. Particularly, the SVM model consistently outperformed across almost all datasets and models. The strong performance of the CropGAN dataset across several models is attributable to the architectural design rather than being generic across other synthetic methods of data generation. By this, the study achieves the purpose of extending the traditional GAN architecture beyond image-based tasks to handling multi-crop recommendation datasets that contain both categorical and numeric variables.

However, the present study did not test CropGAN on other regional datasets, and even as DL architectures like CNN and LSTM were employed for model experimentation, Explainable Artificial Intelligence (XAI) was not integrated. Thus, future studies should examine CropGAN on other regional datasets to assess its performance, scalability and generalizability. Integrating XAI approaches will also improve interpretability and trust among farmers and policymakers, especially for the DL models. Deployment of CropGAN-enhanced crop recommendation systems in mobile applications and IoT-enabled advisory platforms to enable real-world applicability is also encouraged.

Data availability

The dataset used in this study is publicly available on the Zenodo repository and can be accessed via the following DOI:

https://doi.org/10.5281/zenodo.19709807.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

Funding

This research received support from the Petroleum Technology Development Fund (PTDF), Abuja, Nigeria, through the In-country Scholarship Scheme (ISS), with the award reference number: PTDF/ED/ISS/PHD/DKK/0059/23.

References

[1] D. K. Kwaghtyo & C. I. Eke, "Smart farming prediction models for precision agriculture: a comprehensive survey", Artificial Intelligence Review **56** (2023) 5729. https://doi.org/10.1007/s10462-022-10266-6.

[2] J. T. Iorzua, D. K. Kwaghtyo, T. P. Hule & A. T. Ibrahim, "AI-driven approach to crop recommendation: tackling class imbalance and feature selection in precision agriculture", Journal of Future Artificial Intelligence and Technologies **2** (2025) 270. https://doi.org/10.62411/faith.3048-3719-118.

[3] R. Akhter & S. A. Sofi, "Precision agriculture using IoT data analytics and machine learning", Journal of King Saud University–Computer and Information Sciences **34** (2022) 5602. https://doi.org/10.1016/j.jksuci.2021.05.013.

[4] Y. Wang, U. S. R. Dhamodharan, N. Sarwar, F. A. Almalki, Q. H. Naith, R. Sathiyaraj & D. Mohan, "A hybrid approach for rice crop disease detection in agricultural IoT system", Discover Sustainability **5** (2024) 99. https://doi.org/10.1007/s43621-024-00285-4.

[5] A. Iorliam, I. B. Iorliam & S. Bum, "Internet of Things for smart agriculture in Nigeria and Africa: a review", International Journal of Latest Technology in Engineering, Management & Applied Science **10** (2021) 7. https://www.researchgate.net/profile/Aamo-Iorliam/publication/349868793_.

[6] F. O. Isinkaye, M. O. Olusanya & A. A. Akinyelu, "A multi-class hybrid variational autoencoder and vision transformer model for enhanced plant disease identification", Intelligent Systems with Applications **26** (2025) 200490. https://doi.org/10.1016/j.iswa.2025.200490.

[7] D. K. Kwaghtyo, C. I. Eke, J. Abah, T. Moses, J. Agushaka & F. B. Fatokun, *Soil and climate parameters-based crop recommendation model for Yandev, Gboko Local Government Area of Benue State*, 2024 18th International Conference on Ubiquitous Information Management and Communication, Kuala Lumpur, Malaysia, 2024, pp. 1–7. https://doi.org/10.1109/IMCOM60618.2024.10418425.

[8] C. J. Ejiyi, D. Cai, F. O. Eze, M. B. Ejiyi, J. E. Idoko, S. K. Asere & T. U. Ejiyi, "Polynomial-SHAP as a SMOTE alternative in conglomerate neural networks for realistic data augmentation in cardiovascular and breast cancer diagnosis", Journal of Big Data **12** (2025) 97. https://doi.org/10.1186/s40537-025-01152-3.

[9] S. Hong, S. An & J. J. Jeon, "Improving SMOTE via fusing conditional VAE for data-adaptive noise filtering", Applied Intelligence **55** (2025) 841. https://doi.org/10.1007/s10489-025-06692-y.

[10] F. Lygerakis & E. Rueckert, "ED-VAE: entropy decomposition of ELBO in variational autoencoders", arXiv preprint, arXiv:2407.06797 (2024). https://arxiv.org/abs/2407.06797.

[11] D. Anuradha, R. Kuchipudi, B. Ashreetha, J. V. N. Ramesh & A. Rami, "Enhancing agricultural yield forecasting with deep convolutional generative adversarial networks and satellite data", International Journal of Advanced Computer Science and Applications **15** (2024) 661. https://doi.org/10.14569/IJACSA.2024.0150269.

[12] M. Fawakherji, V. Suriani, D. Nardi & D. D. Bloisi, "Shape and style GAN-based multispectral data augmentation for crop/weed segmentation in precision farming", Crop Protection **184** (2024) 106848. https://doi.org/10.1016/j.cropro.2024.106848.

[13] I. S. Rahat, H. Ghosh, S. Dara & S. Kant, "Towards precision agriculture tea leaf disease detection using CNNs and image processing", Scientific Reports **15** (2025) 17571. https://doi.org/10.1038/s41598-025-02378-0.

[14] C. Wang, Y. Xia, L. Xia, Q. Wang & L. Gu, "Dual discriminator GAN-based synthetic crop disease image generation for precise crop disease identification", Plant Methods **21** (2025) 46. https://doi.org/10.1186/s13007-025-01361-0.

[15] S. Sarkar, J. Zhou, A. Scaboo, J. Zhou, N. Aloysius & T. T. Lim, "Assessment of soybean lodging using UAV imagery and machine learning", Plants **12** (2023) 2893. https://doi.org/10.3390/plants12162893.

[16] C. Cuenca-Romero, O. E. Apolo-Apolo, J. N. Rodríguez Vázquez, G. Egea & M. Pérez-Ruiz, "Tackling unbalanced datasets for yellow and brown rust detection in wheat", Frontiers in Plant Science **15** (2024) 1392409. https://doi.org/10.3389/fpls.2024.1392409.

[17] M. K. Senapaty, A. Ray & N. Padhy, "A decision support system for crop recommendation using machine learning classification algorithms", Agriculture **14** (2024) 1256. https://doi.org/10.3390/agriculture14081256.

[18] K. G. Sapkal & A. B. Kadam, *Class balancing for soil data: predictive modeling approach for crop recommendation using machine learning algorithms*, EPJ Web of Conferences **328** (2025) 01026. https://doi.org/10.1051/epjconf/202532801026.

[19] M. Iatrou, C. Karydas, X. Tseni & S. Mourelatos, "Representation learning with a variational autoencoder for predicting nitrogen requirement in rice", Remote Sensing **14** (2022) 5978. https://doi.org/10.3390/rs14235978.

[20] Z. Cao, Y. Ma & Z. Zhang, "Corn yield prediction based on remotely sensed variables using variational autoencoder and multiple instance regression", Geoscience and Remote Sensing Letters. arXiv preprint, arXiv:2211.13286 (2022). https://arxiv.org/abs/2211.13286.

[21] M. A. Razavi, A. P. Nejadhashemi, B. Majidi, H. S. Razavi, J. Kpodo, R. Eeswaran, I. Ciampitti & P. V. Vara Prasad, "Enhancing crop yield prediction in Senegal using advanced machine learning techniques and synthetic data", Artificial Intelligence in Agriculture **14** (2024) 99. https://doi.org/10.1016/j.aiia.2024.11.005.

[22] L. Liu & X. Song, "Accurate crop yield prediction via temporal convolutional network and variational autoencoder", Turkish Journal of Agriculture and Forestry **49** (2025) 612. https://doi.org/10.55730/1300-011X.3290.

[23] C. Karam, M. Awad, Y. Abou Jawdah, N. Ezzeddine & A. Fardoun, "GAN-based semi-automated augmentation online tool for agricultural pest detection: a case study on whiteflies", Frontiers in Plant Science **13** (2022) 813050. https://doi.org/10.3389/fpls.2022.813050.

[24] J. J. Bird, C. M. Barnes, L. J. Manso, A. Ekárt & D. R. Faria, "Fruit quality and defect image classification with conditional GAN data augmentation", Scientia Horticulturae **293** (2022) 110684. https://doi.org/10.1016/j.scienta.2021.110684.

[25] L. Shumilo, A. Okhrimenko, N. Kussul, S. Drozd & O. Shkalikov, "Generative adversarial network augmentation for solving the training-data imbalance problem in crop classification", Remote Sensing Letters **14** (2023) 1129. https://doi.org/10.1080/2150704X.2023.2275551.

[26] Q. B. Kwong, Y. T. Kon, W. R. W. Rusik, M. N. A. Shabudin, S. S. A. Rahman, H. Kulaveerasingam & D. R. Appleton, "Enhancing oil palm segmentation model with GAN-based augmentation", Journal of Big Data **11** (2024) 126. https://doi.org/10.1186/s40537-024-00990-x.

[27] E.-J. Kim & P. Bansal, "A deep generative model for feasible and diverse population synthesis", arXiv preprint, arXiv:2208.01403 (2023). https://doi.org/10.48550/arXiv.2208.01403.

[28] G. Douzas & F. Bacao, "Geometric SMOTE: A Geometric enhanced drop-in replacement for SMOTE", Information Sciences **501** (2019) 118–135. https://doi.org/10.1016/j.ins.2019.06.007.

Appendix A. Samples of synthetic datasets generated using SMOTE and VAE methods.

Table A.1: SMOTE generated sample dataset.

	N	P	K	temp	hum	ph	rainf	soil	label
0	19.0	65.0	25.0	18.10	18.29	5.63	144.79	0.0	0.0
1	36.0	43.0	22.0	27.83	87.17	6.39	58.37	1.0	9.0
2	99.0	5.0	47.0	24.13	84.84	6.65	51.19	2.0	5.0
3	25.0	68.0	77.0	20.09	15.11	7.70	85.75	1.0	7.0
4	58.0	73.0	16.0	33.37	65.68	6.87	64.90	3.0	1.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
9995	21.0	38.0	21.0	29.76	86.45	6.64	37.55	1.0	9.0
9996	22.0	56.0	17.0	29.88	87.33	6.89	44.75	1.0	9.0
9997	8.0	54.0	20.0	28.33	80.77	7.03	38.80	1.0	9.0
9998	4.0	41.0	20.0	28.15	83.80	6.65	37.45	1.0	9.0
9999	27.0	40.0	24.0	27.84	90.00	7.06	52.85	1.0	9.0
10000 rows × 9 columns									

Table A.2: VAE generated sample dataset.

	N	P	K	temp	hum	ph	rainf	soil	label
0	19.00	65.00	25.00	18.10	18.29	5.63	144.79	0.00	0.0
1	36.00	43.00	22.00	27.83	87.17	6.39	58.37	1.00	9.0
2	99.00	5.00	47.00	24.13	84.84	6.65	51.19	2.00	5.0
3	25.00	68.00	77.00	20.09	15.11	7.70	85.75	1.00	7.0
4	58.00	73.00	16.00	33.37	65.68	6.87	64.90	3.00	1.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
9995	24.72	47.21	20.69	28.19	84.25	6.67	53.48	1.03	9.0
9996	21.12	47.63	18.64	28.66	85.34	6.64	46.84	1.03	9.0
9997	18.11	46.35	19.38	28.58	85.89	6.67	45.39	0.94	9.0
9998	26.20	46.71	17.02	28.59	86.22	6.61	44.46	0.97	9.0
9999	20.75	48.03	20.37	28.36	85.90	6.75	47.48	1.01	9.0
10000 rows × 9 columns									