ATTENTION BASED CONDITIONAL GAN FOR SYNTHETIC CROP DATA: SOLVING AGRICULTURE'S DATA AVAILABILITY AND QUALITY CHALLENGES

M.J. Aashik Rasool , Abrar Alabdulwahab , Sevara Mardieva, Dong Seok Kim

ABSTRACT

Agriculture is pivotal for the global economy and sustenance. However, it confronts challenges from a burgeoning population, climate shifts, and the imperative for sustainable practices. Artificial intelligence (AI) based solutions offer promise, but the need for substantial-high-quality training data in agriculture is impractical. Existing methods for generating synthetic data face significant challenges regarding precision and reliability, compromising their effectiveness in complex AI-based models for agriculture. To overcome this, we propose an attention-based conditional Generative Adversarial Network enhanced with correlation coefficients from original datasets. Unlike existing methods, our approach effectively replicates the intricacies of real-world agricultural data. Through comprehensive evaluations, we validate its superior performance in producing realistic and relevant synthetic datasets. Incorporating correlation coefficients as a condition and utilizing multi-head attention in the generator, our approach effectively captures the intricate relationships in agricultural data. Leveraging these data enables the training of more precise and accurate models for the agricultural field. Our code is available at: https://github.com/aashikrasool/Coefficient-Based-Data-Generator

1 Introduction

Agriculture plays a significant role in the global economy and food safety [1],[2]. It contributes to many countries' Gross Domestic Product (GDP), where they depend on agricultural imports and exports for their economic stability. A sustainable economy will create jobs and offer a lifeline for rural areas. Agriculture also plays an essential part in the environmental management of natural resources, such as biodiversity, reforestation, and the sustainable management of natural resources. These Agriculture activities lead to preserving the ecosystem and making agriculture crucial for today's and future generations [3],[4].

Data plays a crucial role in implementing artificial intelligence (AI) techniques in agriculture [5],[6]. The performance and efficiency of training models in AI applications are directly proportional to the size of the dataset collection; a more substantial dataset typically yields better results. However, unlike accessible collected datasets such as Healthcare and financial data, collecting substantial data in agriculture can be impractical. Therefore, researchers propose methods like synthetic data generation to address similar data challenges in many sectors. In Generative Adversarial Network (GAN), researchers actively use synthetic data generation for image and tabular data [7]. GAN architectures typically comprise two parts: a generator that generates synthetic outputs from random noise and a discriminator that distinguishes between the generator's generated data and real data [8]. Synthetic data can converge toward real data quality by coordinating the conditions with the generator. Therefore, deep learning models can be trained by utilizing the generated converged agriculture data.

Initially, Park et al. [9] proposed an approach called "TabGAN" to generate tabular data by capturing the statistical properties of tabular data represented via the GAN approach. Subsequently, Xu et al. [10] proposed an approach called "CTGAN," which specifically addresses the imbalanced data and complex multivariate relationships within categorical data. However, these approaches encounter a substantial challenge in generating precise synthetic data, with the added

[1]All authors contributed equally to this work.

[1]This work was conducted for the IT Convergence Department at Gachon University, South Korea, as part of the Data-Centric Class project during the 2023-2 semester.

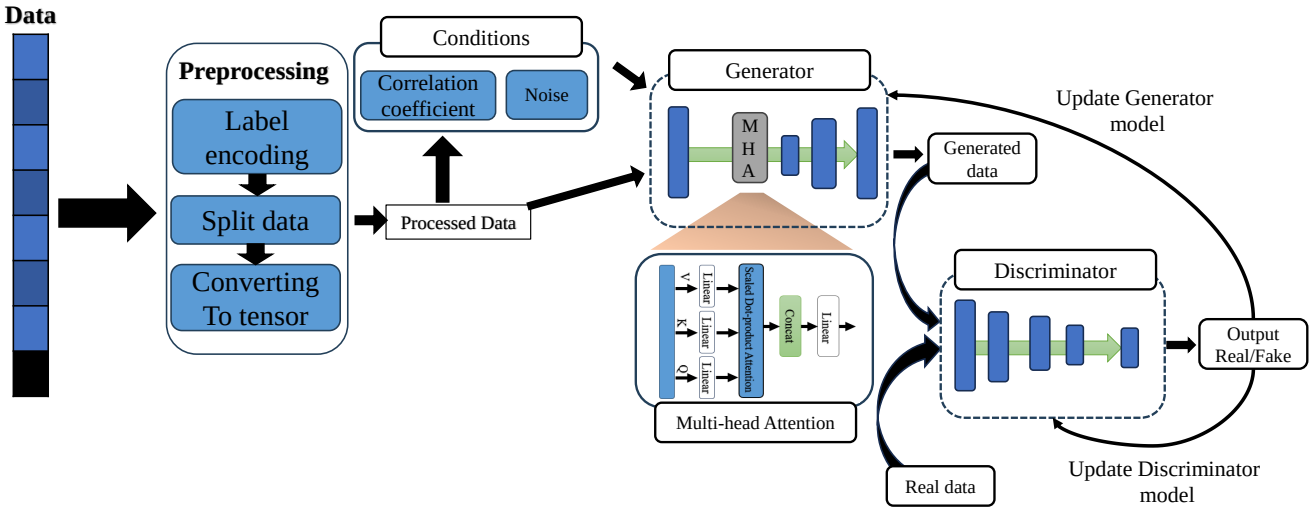


Figure 1: our proposed GAN approach, utilizing a correlation coefficient as a guiding condition for the generator. The generator features a Multi-Head Attention layer, efficiently processing 'Values', 'Keys', and 'Queries' from tabular data for enhanced synthetic data generation.

issue of unevenly distributed data volume among different classes. Some classes have excess data, while others need more data, significantly undermining the performance of sophisticated AI-driven agricultural models. To overcome these challenges, we propose a method to generate a rational and efficient agricultural table dataset using attention-based conditional GAN, aiming to contribute in the following ways:

- Propose an advanced conditional GAN that integrates correlation coefficients as conditional inputs, enhancing the model's ability to mirror complex statistical relationships in agricultural data accurately. This approach offers a novel solution to the limitations of existing tabular data generation methods by generating precise and evenly distributed synthetic tabular data. To the best of our knowledge, this method represents the initial approach to generating agricultural data. It is deemed a more accurate data generation method compared to existing methods.
- Our framework introduces a multi-head attention layer within the generator. This innovation allows for a more nuanced discernment and emphasis on critical inter-feature relationships in tabular data, significantly improving the model's capability to capture and replicate complex data dynamics in the agricultural domain.
- We thoroughly evaluated various state-of-the-art models, including our proposed approach, and analyzed their respective performance measures.

2 Proposed Approach

In contrast, GAN models play a major role in contemporary generative models. In our proposed approach, we utilize conditional GAN to construct tabular data in a precise manner. Conditional GANs [11] is designed to learn a mapping that incorporates both observed data, denoted as $I1$, and a random noise vector along with a correlation coefficient (r) as conditions. This process aims to generate an output represented as $I2$. We can express the mapping as follows:

$$G : \{I1, r\} \rightarrow I2 \tag{1}$$

We train the generator (G) to produce outputs that the discriminator cannot distinguish from real data. This training process uses a correlation coefficient and the original data shape as inputs, acting as noise to guide the generation. A discriminator (D), adversarially trained for this purpose, evaluates the generator's performance. Its primary role is identifying any artificial or fake data the generator produces. This adversarial dynamic ensures continuous generator improvement, leading to increasingly realistic data generation.

2.1 Correlation coefficient as a condition to Generator

In scientific discourse, a correlation coefficient is a key statistical tool that quantifies the extent and manner in which two variables are associated. It ranges from -1 to 1, indicating strong negative to strong positive linear relationships, respectively. Zero implies no linear correlation. Pearson's coefficient is used for continuous, normally distributed data, Spearman's for ordinal or non-normal data, and Kendall's for smaller datasets. In addition, it is crucial in research to choose the appropriate coefficient based on data properties and to remember that correlation does not imply causation. [12]. In the proposed model we use Pearson's correlation coefficient as a condition of the generator along with a data shape as a noise. The reason we selected Pearson's correlation coefficient is that most agricultural data, including [13] data, are continuous and normally distributed.

2.1.1 Pearson's correlation coefficient:

Pearson's correlation coefficient, symbolized as ρ for population and r for sample data, assesses the linear association between two variables that are both normally distributed. This value can be skewed by outliers, which may amplify or diminish the perceived correlation, rendering it less suitable for variables that deviate from normal distribution. The calculation of r for sample data involves a specific formula that considers the relationship between the variables x and y [12].

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

where x_i and y_i are the values of x and y for the i^{th} element.

2.2 Multi-head attention-based Generator

In our proposed model, the generator comprises five layers. Notably, it integrates a multi-head attention mechanism, as described in Pan et al. [14], into its second layer. This essential layer processes three types of input: values (V), keys (K), and queries (Q). The processing involves the calculation of attention weights, formulated as Weights $=$ softmax $\left(\frac{Q_x K_x^T}{\sqrt{d_k}}\right)$. This computation leverages the scaled dot-product attention mechanism, dynamically assigning varying levels of significance to different segments of the data based on V_x, K_x, and Q_x. This strategic allocation enables the model to concentrate on the most relevant information, a key factor in enhancing its capability to synthesize pertinent data. The detailed workings of this method are systematically outlined in Algorithm 1.

Initially, our generator receives the tabular data, correlation coefficient value, and data shape as an input vector (refer to Line 1). In Line 3, multi-head attention plays a crucial role in capturing complex relationships within the data. For each instance x in the input data I_1 (Line 5), the algorithm proceeds to process the inputs V_x, K_x, and Q_x. Subsequently, it computes attention weights to assess the relevance of various elements within the data (Line 6).

Algorithm 1 Multi-head attention-based generator for tabular data generation

1:
Require: I_1 (Tabular Data), r (Correlation Coefficient), S (Data Shape)
Ensure: I_2 (Synthetic Tabular Data)
2: Initialize a 5-layer generator, G
3: Multi-head attention in the second layer of G
4: **for** each instance x in I_1 **do**
5: Process inputs: V_x, K_x, Q_x
6: Weights = softmax($\frac{Q_x K_x^T}{\sqrt{d_k}}$)
7: Assign weights dynamically based on V_x, K_x, Q_x
8: **if** corr(x) $\geq r$ and shape(x) = S **then**
9: Prioritize x for targeted synthesis
10: **end if**
11: **end for**
12: Produce I_2 with properties of I_1 =0

Instances that meet specific criteria, such as a predefined correlation coefficient threshold and adherence to a certain data shape (S), are granted priority for targeted synthesis. This prioritization step occurs between Lines 8 and 9. Finally,

Table 1: Comparative analysis of Correlation Coefficients (CC) relative to original data for proposed Model and existing State-of-the-Art tabular data generation Methods.

Model	CC compared to original data
TabGAN[9]	0.631
CTGAN[10]	0.752
Proposed model	**0.989**

the algorithm generates synthetic tabular data (I_2) with properties akin to those of the input data I_1 (Line 12). This synthesis step guarantees that the generated data preserves the fundamental characteristics of the original dataset.

3 Experiments and Results

This section provides a comprehensive description of the data preprocessing methods and results comparison between state-of-the-art methods.

3.1 Dataset

In this research, we utilized a crop recommendation dataset [13] sourced from Kaggle for our analysis. This data was released by the Indian Chamber of Food and Agriculture. This dataset comprises 22 distinct types of crops, amounting to a total of 2,200 data entries. The dataset is tabular in format, containing seven numerical columns and one categorical column.

3.1.1 Data Preprocessing:

Initially, we transformed the categorical column into a numeric format using the label encoding mechanism. Subsequently, we divided the dataset into training and testing sets, maintaining a ratio of 80% for training and 20% for testing. Afterward, we converted the data into tensors, preparing it to be fed as input into our generator.

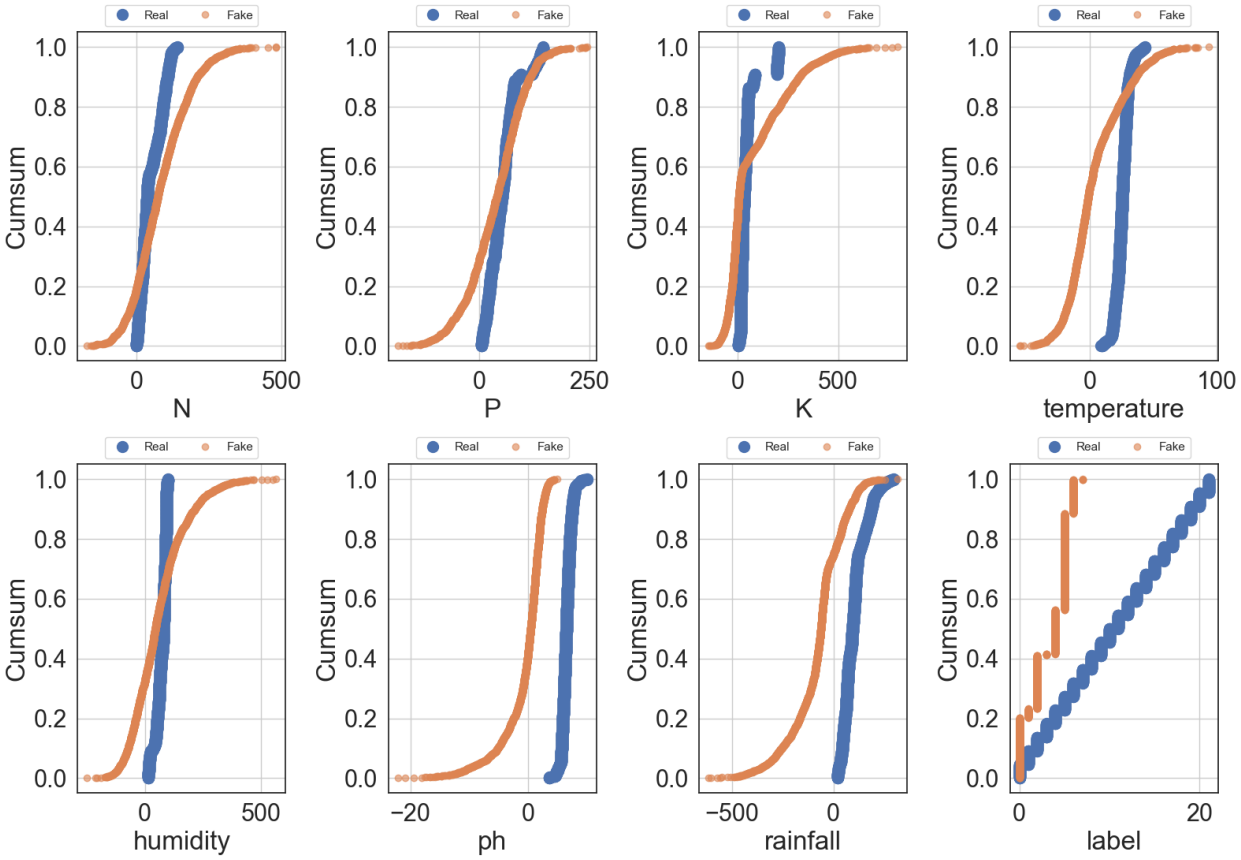
3.2 Comparative results

Table 1 provides a comparison of GAN models tailored for tabular data, employing correlation coefficient (CC) as the metric for evaluation. The CTGAN [10] records a strong CC of 0.752, and TabGAN [9] registers a moderate CC of 0.631. The proposed model demonstrates superior performance with a CC of 0.989, indicative of an almost perfect linear association and thus suggesting a highly effective model in relation to the expected outcomes.

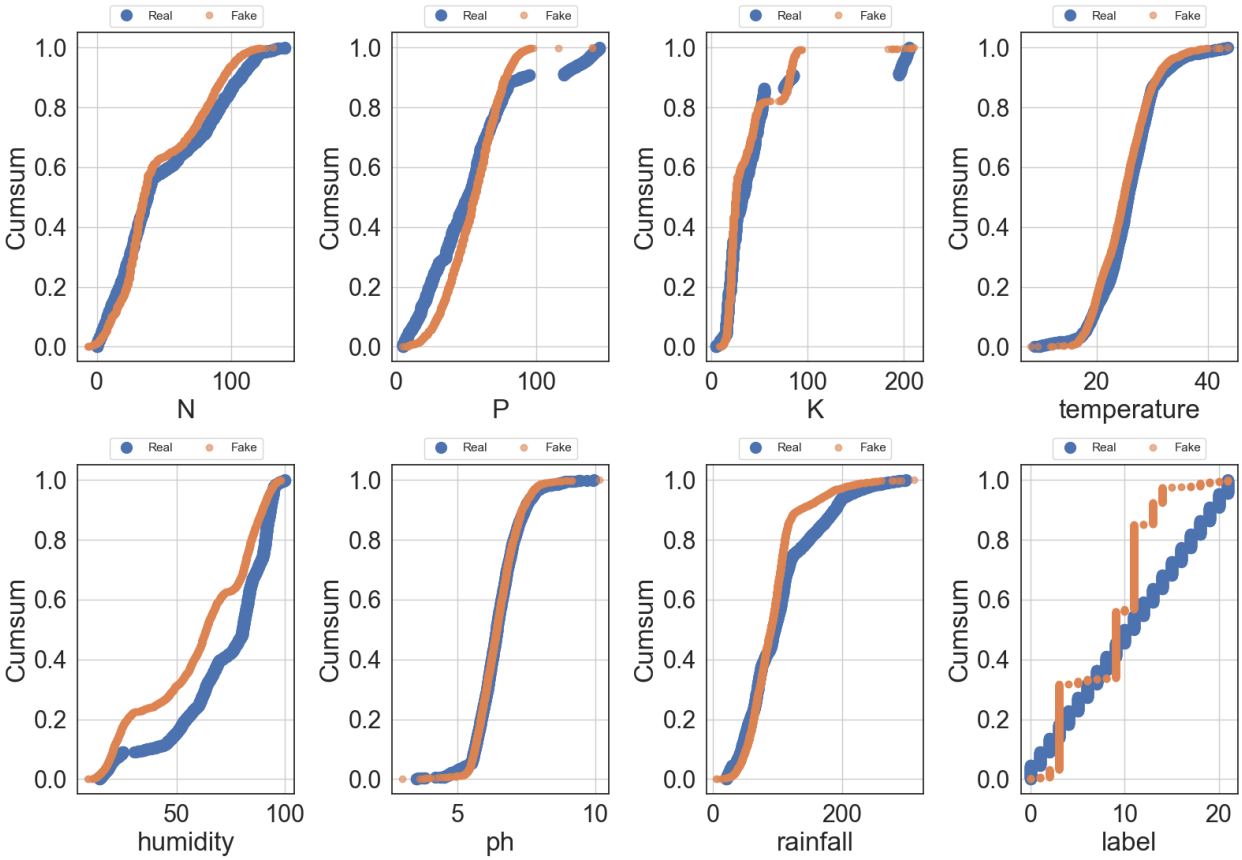
In addition, Figure **??** represents the cumulative sums of the cumulative distribution functions(CDFs) per feature for each of the three methods. Each graphical representation delineates the cumulative probability of 'Real' versus 'Fake' datasets with respect to distinct feature values, as plot along the abscissa. Visualization of the CDFs are instrumental for elucidating both the distributional characteristics and data point variability, facilitating a comprehensive comparative analysis over the complete value range of individual features. Figure **??** shows 8 plots for each model, totally 24. Figure 2a demonstrates CDFs of the TabGAN, where the model generates only 8 labels out of 21 that visible in the 'label' plot (8/8). The label 'N'(Nitrogen) shows a close alignment distribution, with a slight divergence at higher values, while there are significant differences between the synthetic data and the original data for labels 'pH' and 'Rainfall'. In contrast, CTGAN demonstrates more better results by generating data without mislabelling issue, Figure 2b. Though the labels 'pH' and 'Temperature' show almost 95% probability distribution, the number of each label is dramatically various. Moreover, it is the limitation of tabular GANs that, they are not able to generate specific amount of data. Nevertheless, our proposed approach performs better in producing evenly distributed data that closely resembles the original data in all columns with likeness of 98%, except for the 'P'(Phosphorus) column, where the synthetic data shows a more rapid increase initially, suggesting a higher density of lower values compared to real data, Figure 2c. As a reason we can demonstrate several instances, and one of them is; the feature itself might have a natural distribution that is represented in this manner in synthetic data. For the future work, we improve our model for this kind of limitations.

4 Conclusion

The prevalent challenge in agricultural AI has been the need for more high-quality data, which hampers the development and effectiveness of advanced models. Our study addresses this fundamental issue by introducing an innovative

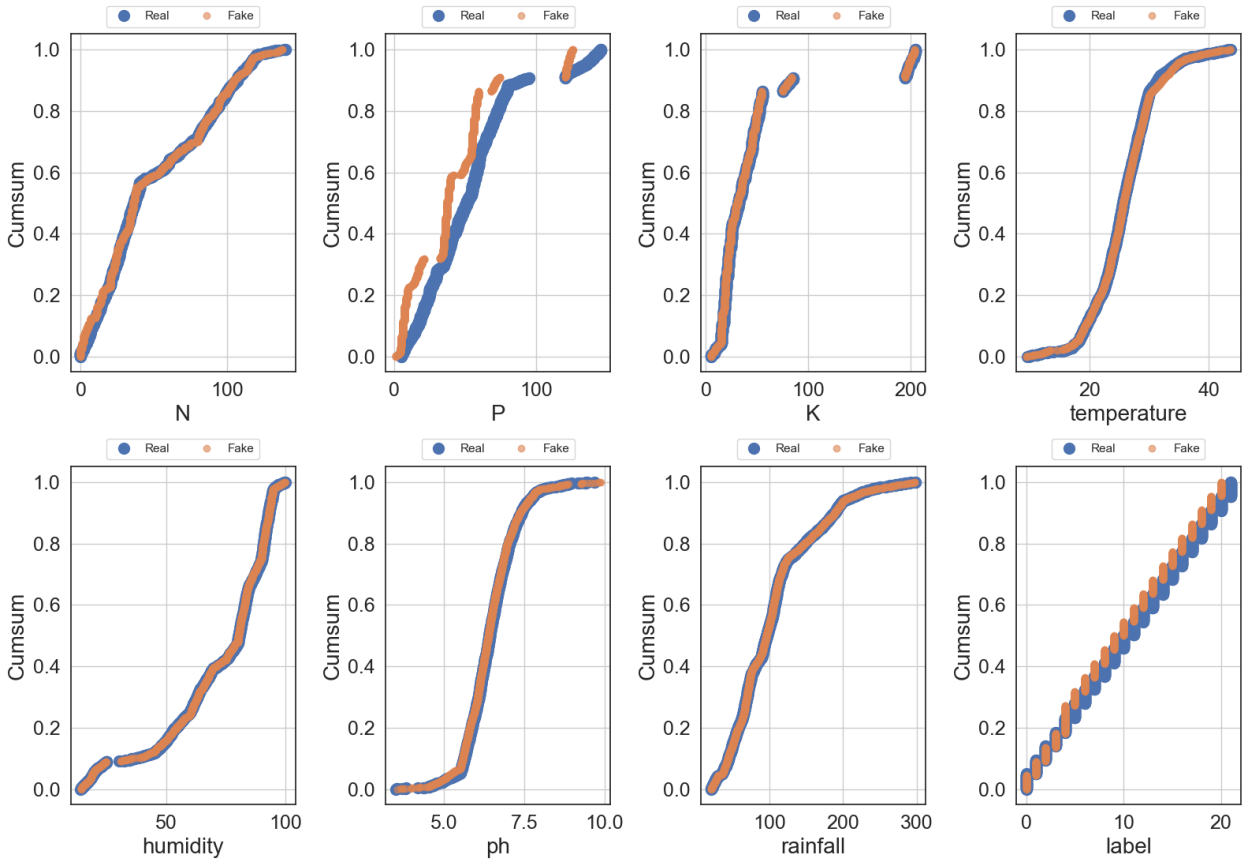


(a) Cumulative sums per features for TabGAN



(b) Cumulative sums per features for CTGAN

Figure 2: Comprehensive analysis of cumulative sums per features (Part 1).



(c) Cumulative sums per features for the proposed model

Figure 2: Comprehensive analysis of cumulative sums per features (Part 2).

attention-based conditional GAN augmented with correlation coefficients from original datasets. This approach marks a significant improvement over existing methods for synthetic data generation, which have struggled with precision and reliability issues. Our model excels in accurately replicating the complex patterns of real-world agricultural data, thereby generating synthetic datasets that are both realistic and relevant. The comprehensive evaluations of our model confirm its superior performance, highlighting its capability to overcome the limitations of data scarcity in agricultural AI. This research provides a novel solution to a critical problem and paves the way for future advancements in AI-driven agricultural technologies.

References

[1] Lontsi Saadio Cedric, Wilfried Yves Hamilton Adoni, Rubby Aworka, Jérémie Thouakesseh Zoueu, Franck Kalala Mutombo, Moez Krichen, and Charles Lebon Mberi Kimpolo. Crops yield prediction based on machine learning models: Case of west african countries. *Smart Agricultural Technology*, 2:100049, 2022.

[2] Getachew Bezabih, Melaku Wale, Neela Satheesh, Solomon Workneh Fanta, and Minaleshewa Atlabachew. Forecasting cereal crops production using time series analysis in ethiopia. *Journal of the Saudi Society of Agricultural Sciences*, 2023.

[3] K Arumugam, Yarnagula Swathi, Domenic T Sanchez, Malik Mustafa, Chirasak Phoemchalard, Khongdet Phasinam, and Ethelbert Okoronkwo. Towards applicability of machine learning techniques in agriculture and energy sector. *Materials Today: Proceedings*, 51:2260–2263, 2022.

[4] Vandana Kushwaha, Priya Shukla, and Gora Chand Nandi. Generating quality grasp rectangle using pix2pix gan for intelligent robot grasping. *Machine Vision and Applications*, 34(1):15, 2023.

[5] Mohd Javaid, Abid Haleem, Ibrahim Haleem Khan, and Rajiv Suman. Understanding the potential applications of artificial intelligence in agriculture sector. *Advanced Agrochem*, 2(1):15–30, 2023.

[6] Ashok Kumar Koshariya, D Kalaiyarasi, A Arokiaraj Jovith, T Sivakami, Dler Salih Hasan, and Sampath Boopathi. Ai-enabled iot and wsn-integrated smart agriculture system. In *Artificial Intelligence Tools and Technologies for Smart Farming and Agriculture Practices*, pages 200–218. IGI Global, 2023.

[7] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493:28–45, 2022.

[8] Abeer Aljohani and Nawaf Alharbe. Generating synthetic images for healthcare with novel deep pix2pix gan. *Electronics*, 11(21):3470, 2022.

[9] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *arXiv preprint arXiv:1806.03384*, 2018.

[10] L Xu, M Skoularidou, A Cuesta-Infante, and K Veeramachaneni. Modeling tabular data using conditional gan. arxiv 2019. *arXiv preprint arXiv:1907.00503*, 1, 1907.

[11] Jordan J Bird, Chloe M Barnes, Luis J Manso, Anikó Ekárt, and Diego R Faria. Fruit quality and defect image classification with conditional gan data augmentation. *Scientia Horticulturae*, 293:110684, 2022.

[12] Mavuto M Mukaka. A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal*, 24(3):69–71, 2012.

[13] Siddharth S. Crop recommendation dataset. `https://www.kaggle.com/datasets/siddharthss/crop-recommendation-dataset`, 2023. Accessed: 10/10,.

[14] Tongyang Pan, Jinglong Chen, Zhisheng Ye, and Aimin Li. A multi-head attention network with adaptive meta-transfer learning for rul prediction of rocket engines. *Reliability Engineering & System Safety*, 225:108610, 2022.