Engineering Applications of Artificial Intelligence 131 (2024) 107881

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

A comprehensive review of synthetic data generation in smart farming by using variational autoencoder and generative adversarial network

Yaganteeswarudu Akkem [a,*], Saroj Kumar Biswas [a], Aruna Varanasi [b]

[a] Computer Science & Engineering, National Institute of Technology, Silchar, Cachar, Assam, India
[b] Computer Science & Engineering, Sreenidhi Institute of Science & Technology, Yamnampet, Ghatkesar, Hyderabad, Telangana, India

ARTICLE INFO

Keywords:
Variational autoencoders
Generative adversarial networks
Smart farming

ABSTRACT

In this study, we propose the use of Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) to generate synthetic data for crop recommendation (CR). CR is critical in agriculture, assisting farmers in making informed decisions about crop cultivation, considering factors like soil conditions, weather patterns etc. Unfortunately, the availability of labeled data for CR is often limited, posing a significant challenge in training accurate recommendation models. VAEs and GANs are employed to create synthetic data that closely mirrors real-world crop data. VAEs are utilized to extract latent representation from the input data, enabling the generation of new samples with similar characteristics. GANs play a crucial role in generating data by training a generator network to produce synthetic samples that closely resemble real data, while a discriminator network distinguishes between genuine and synthetic data. The generated synthetic data serves as a valuable resource to prepare datasets for CR, enhancing the performance of recommendation models. Our research explores the effectiveness of VAEs and GANs in producing high-quality synthetic CR data, facilitating improved training and evaluation of recommendation systems. This paper presents the architecture and training process of the proposed models and evaluates the quality and utility of the generated synthetic data using various experiments, including visualizations such as heatmaps, scatter plots, cumulative sum per feature plots, and distribution per feature plots. The results of this study hold the potential to make a significant contribution to the field of agriculture by providing a reliable and abundant source of training data for CR systems.

1. Introduction

The field of agriculture heavily relies on crop recommendation systems to guide farmers on the most suitable crops to cultivate, considering various factors such as soil conditions, weather patterns, and market demand. However, the limited availability of labeled data for crop recommendation poses a significant challenge in training accurate models. This research proposes a novel approach to this problem by leveraging artificial intelligence techniques, specifically VAEs, GANs, and language models, to generate synthetic data that closely mirrors real-world crop data. The use of VAEs allows for the learning of a latent representation of the input data, enabling the generation of new samples with similar characteristics. GANs, on the other hand, are employed to generate data that is hard to distinguish from real ones. This study aims to explore the effectiveness of these techniques in generating high-quality synthetic crop recommendation data, thereby enhancing the training and evaluation of recommendation systems. The proposed models' architecture and training procedure will be covered in full in this study, along with an assessment of the accuracy and value of the generated synthetic data. The findings of this research could significantly contribute to the field of agriculture by providing a reliable and abundant source of training data for crop recommendation systems.

Introducing synthetic data generation techniques could revolutionize how we approach data scarcity in the agricultural sector. By creating high-quality, contextually accurate synthetic data, we can overcome the limitations of existing datasets and improve the accuracy of crop recommendation models. This study will provide an in-depth analysis of the design, training procedure, and general efficacy of VAEs and GANs for the generation of synthetic data. We will also present a thorough evaluation of the generated synthetic data, using various metrics and experiments to assess its quality and usefulness. The potential implications of this research are vast, offering a promising solution to data scarcity and paving the way for more advanced, accurate crop recommendation systems. This could significantly enhance

* Corresponding author.
E-mail address: yaganteeswarudu21_rs@cse.nits.ac.in (Y. Akkem).

https://doi.org/10.1016/j.engappai.2024.107881
Received 30 September 2023; Received in revised form 12 December 2023; Accepted 8 January 2024
Available online 19 January 2024
0952-1976/© 2024 Elsevier Ltd. All rights reserved.

Table 1
Literature review for the importance of Generative AI in smart farming.

Author(s)	Title	Research Objectives	Importance of Generative AI in Smart Farming
Chia, Y.K. et al. (2022)	Relationprompt: Leveraging prompts.	Create fictitious data for zero-shot triplet relation extraction.	Generative AI can create synthetic data to improve machine learning models used in smart farming, enhancing decision-making and prediction accuracy.
Dai, H et al. (2023)	Chataug: Leveraging chatgpt for text data augmentation	Utilize ChatGPT for text data augmentation	ChatGPT-based text data augmentation can enhance the quality and quantity of textual data in agriculture-specific applications, improving insights and decision support.
Jain et al. (2019)	Agribot: agriculture-specific question-answer system	Develop an agriculture-specific question-answering system	AI-driven question-answering systems can provide farmers with immediate answers to agriculture-related queries, promoting informed decision-making.
Li, X. et al. (2021)	Weather GAN: Multi-domain weather translation	Translate weather information across multiple domains using GANs	Generative AI, like GANs, can facilitate the translation of weather data across different domains, aiding precision agriculture and crop management.
Liu, Y et al. (2023)	Summary of chatgpt/gpt-4 research.	Provide a summary of ChatGPT/ GPT-4 research.	ChatGPT/GPT-4 and similar models can play a crucial role in advancing AI applications in agriculture, offering new possibilities for automation and data analysis.
Rezayi, S. et al. (2022)	Agribert: knowledge-infused agricultural language models	Develop knowledge-infused language models for food and nutrition matching	Knowledge-infused language models like AgriBERT can help in matching food and nutrition data, aiding in crop planning and dietary recommendations in smart farming.
Agarwal, O. et al. (2021)	Knowledge graph-based synthetic corpus generation.	Generate synthetic corpora for knowledge-enhanced language model pre-training	Synthetic corpora generation can enhance the training of knowledge-enhanced language models, which can be applied in various aspects of smart farming.
Xu Han (2018)	Fewrel: A large-scale supervised few-shot relation classification.	Create a large-scale dataset for few-shot relation classification	Large-scale datasets like FewRel can be used to train AI models for agriculture-specific tasks, such as relation classification in farm data.
Keskar, (2019)	Ctrl: A conditional transformer language model for controllable generation.	Develop a conditional transformer language model for controllable text generation.	Controllable text generation models like Ctrl can assist in generating customized agricultural reports, summaries, or explanations, enhancing user interaction in smart farming.
Pengfei Liu et al. (2021)	Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing.	Conduct a survey of prompting methods in NLP.	Knowledge of effective prompting methods can be applied to develop user-friendly interfaces for smart farming applications, facilitating human-AI interaction.
Yu Meng et al. (2022)	Generating training data with language models: Towards zero-shot language understanding.	Explore generating training data for zero-shot language understanding tasks.	Zero-shot language understanding can enable AI systems to adapt quickly to new agricultural tasks and information, improving their versatility.
Laria Reynolds and McDonell (2021)	Prompt programming for large language models.	Investigate prompt programming techniques for large language models	Advanced prompt programming can enable large language models to perform complex tasks specific to smart farming, such as data analysis and decision support.
Jiacheng Ye et al. (2022)	Zerogen: Efficient zero-shot learning via dataset generation.	Develop efficient methods for zero-shot learning via dataset generation	Efficient zero-shot learning can be applied in agriculture to adapt AI models to new scenarios without extensive manual data labeling, improving automation.

agricultural decision-making processes, leading to increased productivity and sustainability.

Smart farming Akkem et al., 2023a,2023b, also known as precision agriculture, integrates technology and data-driven techniques to optimize agricultural practices to optimize resource utilization, improve crop yield, and reduce environmental impact. Due to its capacity to analyze massive volumes of data and make wise judgments, AI is being used in smart farming. Generative AI, a subset of AI that focuses on generating new data instances that resemble a given dataset, has shown promise in various domains. This review aims to explore the applications, benefits, and challenges of using generative AI techniques in the context of smart farming.

The use of Generative AI in smart farming has opened up innovative opportunities to address various agricultural complexities. Smart farming has emerged as a critical approach to addressing the pressing challenges of our time, including the increasing global demand for food, the impacts of climate change, and the scarcity of vital agricultural resources. In this context, Generative Artificial Intelligence (AI) techniques, notably GANs and VAEs, have emerged as powerful tools with significant potential to revolutionize various aspects of smart farming. The ultimate goal of this research is to contribute to the advancement of agricultural practices by providing a reliable and abundant source of training data for crop recommendation systems. By harnessing the power of artificial intelligence and synthetic data, we aim to enhance the accuracy and efficiency of these systems, thereby aiding farmers in making informed decisions about crop cultivation. The findings of this research could have far-reaching implications, not only for the field of agriculture but also for other sectors facing similar challenges with data scarcity.

Accurate crop recommendations can potentially revolutionize modern agriculture by addressing the challenges farmers face. Consider a real-world scenario where a farmer situated in a region with varying soil types and weather conditions needs to decide which crops to cultivate for the upcoming season. The decision involves optimizing factors like soil conditions, climate patterns, and market demand. In the absence of comprehensive, labeled data, traditional recommendation systems may fall short in providing tailored advice.

Our research, which leverages VAEs and GANs to generate synthetic crop data, can directly impact this scenario. By using synthetic data that closely mirrors the real-world conditions of the farmer's region, we can create more accurate and personalized crop recommendations. This means the farmer can make data-driven decisions, leading to improved crop yields, reduced resource wastage, and increased profitability.

Furthermore, consider a case study involving a region facing a sudden shift in weather patterns and climate change. Accurate and adaptable crop recommendations become even more crucial in such circumstances. By integrating user feedback and expert knowledge into the data generation process, we can continuously refine the recommendation models to account for changing conditions, thereby helping farmers navigate the challenges posed by climate variability.

2. Literature review

Table 1 represents various articles referred to study the importance of generative AI in crop recommendation. Smart farming, characterized by integrating advanced technologies into agricultural practices, has become pivotal in addressing the challenges posed by increasing global food demand, climate change, and resource scarcity. In this context,

Generative AI techniques, including GANs and VAEs, have emerged as powerful tools with the potential to revolutionize various facets of smart farming. The applications of generative AI in the field of smart farming are thoroughly explored in this literature review, with an emphasis on their contributions to crop suggestions. Accurate yield prediction is a cornerstone of efficient agricultural practices. Li et al., (2021) presented "Weather GAN," a multi-domain weather translation model based on GANs. This technology enables farmers to make data-driven decisions by translating weather data from various sources and domains, facilitating more accurate yield predictions and better resource allocation. Effective disease detection is vital for ensuring crop health and maximizing yields. The study by Rezayi et al., (2022) introduced Agribert, knowledge-infused agricultural language models, which can significantly contribute to disease detection by leveraging domain-specific knowledge. These models can analyze textual and image data related to diseases, offering early Detection and precise recommendations for disease control measures. Generative AI, particularly GANs, has demonstrated its prowess in optimizing crop management practices. Chia et al., (2022) leveraged generative models to synthesize synthetic data for zero-shot relation triplet extraction, which can potentially enhance data-driven decision-making in crop management. The ability to generate synthetic data can aid in creating more robust models for crop yield prediction, disease detection, and pest control, thereby improving crop management efficiency. Resource optimization is crucial for sustainable farming practices. Agarwal et al., (2021) introduced a knowledge graph-based synthetic corpus generation method that enhances language model pre-training. This approach can help optimize resource allocation and decision-making in smart farming systems by generating synthetic data based on agricultural knowledge graphs. Despite the promise of generative AI in smart farming, several challenges persist. These include the need for larger and more diverse datasets, model interpretability, and addressing ethical considerations related to data usage by Liu et al., (2021). Additionally, Jangir et al. (2023) introduced ensuring the scalability and accessibility of these technologies to farmers with varying levels of technical expertise is crucial.

Table 2 represents various articles referred from various journals to study how VAEs work to generate synthetic data. VAEs are capable of learning a latent space representation of data, allowing them to capture essential features and patterns in the input data. In the context of crop recommendation, VAEs can learn meaningful representations of various crop-related factors, such as soil quality, climate conditions, and historical yield data. VAEs are generative models that can generate new data samples that follow the learned data distribution. In crop recommendation, this ability can be leveraged to create synthetic datasets that simulate various crop growth scenarios and conditions. These synthetic datasets can be valuable for training and evaluating recommendation models. VAEs have been used for anomaly detection tasks in various domains. In the context of smart farming, VAEs can help identify unusual or anomalous conditions in crop fields, such as diseases or pests, by detecting deviations from the learned data distribution. VAEs can be employed in multi-task learning scenarios, where they simultaneously learn to generate data and perform specific tasks related to crop recommendation, such as yield prediction or disease detection. This multi-task learning approach can improve the overall performance of recommendation systems. VAEs can automatically extract relevant features from the input data, reducing the dimensionality of the dataset while preserving critical information. This feature extraction capability can simplify the crop recommendation process and enhance the efficiency of recommendation algorithms. "Green AI Perspective." The effectiveness of VAEs and their capacity to produce synthetic data can support sustainable agriculture by maximizing the use of resources, cutting waste, and minimizing the negative effects of crop production on the environment. While not directly related to crop recommendation, the literature mentions VAEs in the context of recommender systems. VAEs can be adapted to build recommendation models for suggesting crop varieties or agricultural practices based on historical data and user preferences.

Table 3 represents various journal articles studied to find the importance of GANs in synthetic data generation. GANs play a significant role in addressing data scarcity issues in the context of crop recommendation. They can generate synthetic agricultural data, such as crop yield, weather patterns, soil quality, and crop disease data. This synthetic data can complement real-world datasets, making it possible to train more accurate and robust crop recommendation models. GANs are used for data augmentation, particularly in scenarios where the available real-world data is limited. By generating additional synthetic data, GANs help enhance the diversity and volume of training data, which can lead to more reliable crop recommendations. In healthcare applications related to crop farming, GANs can generate synthetic data that preserves the privacy of sensitive information while still allowing for meaningful analysis. This is crucial when dealing with data from individual farms or agricultural practices. Synthetic data generated by GANs can be used to train machine learning models for crop recommendation. These models can learn from the synthetic data to make predictions about optimal crop choices, planting strategies, and crop health management. GANs offer a solution to overcome limitations in the availability of historical agricultural data. They can simulate various agricultural scenarios, including different crop types, environmental conditions, and farming practices, which can be valuable for building versatile crop recommendation systems. Some of the citations mention conditional GANs, which are capable of generating data based on specific conditions or input parameters. This feature can be used to simulate the impact of various factors like weather changes, soil conditions, and pest infestations on crop yields, aiding in more precise recommendations. While GANs are important tools for synthetic data generation in agriculture, some of the citations also discuss challenges and future directions in GANs research. Addressing challenges such as data quality, model stability, and interpretability remains important for advancing the field.

3. Generative AI techniques

3.1. Generative AI techniques and general architectures

Fig. 1 represents GANs General architecture, and GANs are a class of machine learning models designed for generating new data samples that are similar to a given dataset. GANs consist of two neural networks: a generator and a discriminator. These networks are trained in a competitive manner. The generator network takes random noise as input and attempts to generate data samples (e.g., images) that resemble the real data from the training set. The discriminator network tries to distinguish between real data samples from the training set and fake data generated by the generator. The generator and discriminator are trained iteratively. The generator aims to improve its ability to produce realistic data to fool the discriminator, while the discriminator aims to improve its ability to distinguish real from fake data. This adversarial training process results in the generator producing increasingly realistic data samples.

Fig. 2 represents VAEs architecture, and VAEs are generative models that focus on probabilistic data encoding and decoding. They consist of two main components: an encoder and a decoder. The encoder maps data samples to a probabilistic latent space, which means it generates a distribution of possible representations for each data point. The decoder takes a sample from the latent space and generates data that is similar to the input data. VAEs aim to find a latent space where similar data points are close together. Unlike GANs, VAEs generate data probabilistically, allowing for sampling from the latent space to generate diverse outputs. VAEs are used for image generation, data denoising, anomaly detection, and generative modeling tasks. They are suitable when there's a need to explore and manipulate the latent space of data. Table 4 represents a comparison of VAEs and GANs with different features, like how both

Table 2
VAEs literature review.

Author & Year	Title	Research Objective	Advantages	Importance of VAE in Smart Farming
Chen S, Guo W (2023)	Auto-Encoders in Deep Learning	Review of Auto-Encoders in Deep Learning	New Perspectives on Auto-Encoders	High importance in Deep Learning and by using this review, finding basic architecture to produce synthetic data for crop recommendation
Asperti, A., Evangelista, D. (2021)	A Survey on VAE from a Green AI Perspective	Survey on Variational Autoencoders	Green AI perspective on variational autoencoders	Green AI applications include efficient use of land, energy, and water resources with precision agriculture powered by AI.
Lu, G.; Zhao, X.; Yin, J.; (2020)	Multi-task learning using VAEs for sentiment classification	Sentiment classification using VAEs	Multi-task learning for sentiment classification using VAEs	Sentiment analysis in Agriculture can be performed, like how to improve the quality of crops from user feedback.
Diallo, B.; Hu, J.; Li, T.; (2021)	Deep embedding clustering based on contractive Autoencoder	Clustering using contractive Autoencoder	Deep embedding clustering with contractive Autoencoder	Data clustering in Smart Farming can be used to prepare crop recommendation data from unlabelled data
Xie, T.; Fu, X.; Ganea, O.E.; (2021)	Crystal diffusion VAE for periodic material generation	Material generation using VAEs	Crystal diffusion Variational Autoencoder for periodic material generation	Material synthesis in Agriculture, like the creation or production of specific materials that are used to enhance agricultural practices
Andresini, G.; Appice, A.; (2021)	Autoencoder-based deep metric learning for network intrusion detection	Network intrusion detection using Autoencoder	Autoencoder-based deep metric learning for network intrusion detection	How to improve Network security in Smart Farming by using VAEs
Hammouche, R.; Attia, A. (2022)	Gabor filter bank with deep Autoencoder-based face recognition system	Face recognition using deep Autoencoder	Face recognition system using Gabor filter bank with deep Autoencoder	crop type recognition in Agriculture or, disease identification, and so on.
Fan, Y.; Wen, G.; Li, D.; (2020)	Video anomaly detection and localization via VAEs	Video anomaly detection using VAEs	Video anomaly detection and localization using Gaussian mixture FC-VAE	Anomaly detection in Agriculture, like Detection of diseases, pests, and nutrient deficiencies in crops.
Zhang et al. (2020)	A survey of autoencoder-based recommender systems	Survey on autoencoder-based recommender systems	Survey of autoencoder-based recommender systems	Recommender systems in Agriculture to recommend suitable crops to the farmer
Asperti, A. (2019)	About generative aspects of VAEs	Generative Aspects of VAEs	Generative aspects of VAEs	Generative modeling in Smart Farming for synthetic data generation by using VAEs

architectures are scalable, what data types are used, and with many other features.

3.2. Potential strengths of VAEs and GANs for generating synthetic data

VAEs and GANs models are especially suited for synthetic data generation tasks compared to other potential techniques for the following reasons:

VAEs are especially suited for learning a latent representation of the input data because of their unique training method. During the training of VAEs, the model is encouraged to use all the dimensions of the latent space rather than over-relying on only a few. This results in a more effective and diverse use of the latent space, allowing VAEs to represent a wider range of variations in the input data. Also, because VAEs are generative models, they can generate new data points in the data domain after being trained by sampling from the latent space. Another advantage of VAEs is the incorporation of an explicit regularizer in their objective function, which forces the model to learn independent statistical factors in the form of a unit Gaussian. Hence, the learned representation is usually better structured (i.e., more disentangled) and more interpretable compared to some other techniques.

GANs are extremely powerful for generating synthetic data. They achieve this by setting up a game between two neural networks, a generator and a discriminator. The generator tries to produce data that come from some probability distribution. The discriminator then tries to distinguish between samples from the 'real' dataset and synthetic samples from the generator. The generator improves over time to produce synthetic data that are increasingly indistinguishable from the real data. GANs are particularly effective at capturing and reproducing the complex, high-dimensional distributions that describe real-world data better than classical methods like mixture models or sampling techniques. This makes them ideally suited to tasks like image generation, where pixel-level dependencies have to be captured accurately. This powerful feature of GANs comes with a trade-off: they can be notoriously difficult to train, with problems like mode collapse (where the generator generates only a subset of plausible outputs) and instability. There are various modifications of the basic GANs model and training procedure that help overcome these issues to a large extent. Another advantage of GANs is that they can generate very sharp, high-quality samples, which distinguish them from other generative models such as VAEs, where the explicit regularization term can often lead to blurrier samples.

In short, both VAEs and GANs have their strengths and are more suitable for different tasks depending on the application and requirements. They have both brought new capabilities to machine learning, enabling models to understand, generate, and even imagine data in ways that were not possible before.

3.3. Technical depth: Technical details comparisons for the VAE and GAN with respect to agricultural data

In the context of agricultural data, VAEs can be used to capture the complex relationships between various factors such as soil quality, weather conditions, and crop yield. By learning a latent space representation, VAEs can interpolate between data points to generate new samples that are not present in the original dataset but are plausible given the learned data distribution.

Technical Aspects of VAEs for Agricultural Data:

- Latent Space Learning: VAEs learn a continuous, probabilistic latent space representation, which is beneficial for capturing the underlying structure of agricultural data.
- Data Generation: After training, VAEs can generate new data points by sampling from the latent space. This is particularly useful for augmenting agricultural datasets where certain types of data may be scarce.
- Anomaly Detection: VAEs can be used to detect anomalies in agricultural data, such as identifying diseased crops or poor soil conditions, by measuring how well new data fits within the learned distribution.
- Feature Extraction: The encoder part of a VAE can serve as a feature extractor, reducing the dimensionality of the data while retaining important information, which can simplify subsequent analysis.

Table 3
GANs literature review.

Author & Year	Title	Research Objective	Importance of GANs in Smart Farming
Figueira and Vaz (2022)	Survey on Synthetic Data Generation, Evaluation Methods, and GANs	Synthetic data generation and evaluation, GANs are discussed as a part of the survey.	GANs can be used for synthetic data generation of crop recommendation
Goodfellow, I. et al. (2020)	GAN	Introduction to GANs	The basics of GAN will be useful for while synthetic data generation
Emam, KE, et al. (2020)	Introducing Synthetic Data Generation	Introduction to Synthetic Data Generation	The basics of GAN will be useful for while synthetic data generation
Siddani, B. et al. (2021)	Machine learning for physics	Generation of dispersed multiphase flow data	Application in physics-informed data generation
Coutinho-Almeida, J. et al. (2021)	GANs for Tabular Healthcare Data Generation	Healthcare data generation with GANs, Privacy, and utility concerns are discussed.	GANs can be used to preserve privacy in crop recommendation datasets.
Frid-Adar (2018)	Synthetic data augmentation using GAN for improved liver lesion	Data augmentation using GANs for classification and improved liver lesion classification is mentioned.	With limited crop recommendation dataset generating multiple sets by using augmentation
Pan, Z. et al. (2019)	Recent progress on GANs: A survey	A survey of recent progress in GANs Provides an overview of GAN advancements.	The basics of GAN will be useful for while synthetic data generation.
Saxena and Cao (2021)	GANs challenges, solutions, and future directions	Challenges and future directions of GANs, Discusses challenges and potential solutions.	Challenges can be taken into consideration during synthetic crop data generation.
Kim and Myung (2018)	Autoencoder-combined GAN for synthetic image data generation and Detection of jellyfish swarm	Image data generation and Detection Combines autoencoders and GANs for image data gen	Application in image data generation and object detection in smart farming
Xu and Veeramachaneni (2018)	Synthesizing tabular data using generative adversarial networks	Synthesizing tabular data using GANs, Application of GANs in tabular data synthesis	generating synthetic tabular data in smart farming
Xu (2019)	Modeling tabular data using conditional gan	Modeling tabular data using conditional GANs	conditional tabular data modeling in smart farming
Park (2018)	Data synthesis based on GAN	Data synthesis using GANs	Application in data synthesis using GANs and the Basics of GAN will be useful for while synthetic data generation.

- Interpolation: VAEs can interpolate between different data points in the latent space, which can be used to simulate gradual changes in environmental conditions and their impact on crops.

For agricultural data, GANs can generate high-quality synthetic samples that can be used to train crop recommendation systems. This is particularly valuable in scenarios where real-world data is limited or imbalanced.

Technical Aspects of GANs for Agricultural Data:

- Data Augmentation: GANs can augment agricultural datasets by generating new samples that increase the diversity and volume of the data, leading to more robust crop recommendation models.
- High-Quality Generation: GANs are known for their ability to generate sharp and realistic samples, which is crucial for creating synthetic agricultural data that closely resembles real-world conditions.
- Conditional Generation: Conditional GANs can generate data based on specific conditions or parameters, such as simulating crop yields under different weather scenarios.
- Privacy Preservation: GANs can be used to generate synthetic data that maintains the statistical properties of the original data while preserving the privacy of sensitive information.
- Challenges in Training: GANs are notoriously difficult to train, with issues such as mode collapse and training instability. However, various techniques and modifications to the basic architecture can mitigate these problems.

3.3.1. Comparative Analysis

When comparing VAEs and GANs for agricultural data generation, several factors should be considered:

- Quality of Generated Data: GANs often produce more realistic and high-resolution samples compared to VAEs. However, VAEs provide a structured latent space that can be useful for understanding and manipulating the data.
- Training Stability: VAEs generally have a more stable training process due to their probabilistic nature, while GANs require careful tuning to avoid training issues.
- Computational Resources: GANs, especially when generating high-resolution data, can be more computationally intensive than VAEs.
- Interpretability: VAEs offer some level of interpretability through their latent space, while GANs are less interpretable due to their adversarial nature.

3.4. Algorithm complexity and scalability comparisons for the VAE and GAN with respect to agricultural synthetic data

3.4.1. Algorithm complexity and scalability comparison

3.4.1.1. Variational Autoencoders (VAEs).

- Algorithm Complexity:

Variational Autoencoders (VAEs) are a type of generative model that use an encoder-decoder architecture. The encoder transforms the input data into a latent space representation while the decoder reconstructs the input data from this latent space. The complexity of VAEs primarily depends on the architecture of the neural networks used for the encoder and decoder. Typically, these networks are composed of fully connected layers or convolutional layers for image data.

The computational complexity of VAEs is O(n*dˆ2 + d*kˆ2), where n is the number of data points, d is the dimensionality of the input data, and k is the dimensionality of the latent space. The first term (n*dˆ2) corresponds to the encoding process, while the second term (d*kˆ2) corresponds to the decoding process.

- Scalability:

VAEs are moderately scalable to large datasets and high-dimensional data. They can be trained in a mini-batch fashion, which allows them to handle large datasets that do not fit into memory. However, as the dimensionality of the data or the complexity of the network architecture increases, the computational resources required to train VAEs also increase. VAEs can be parallelized across multiple GPUs to improve training time, but this requires significant computational resources.

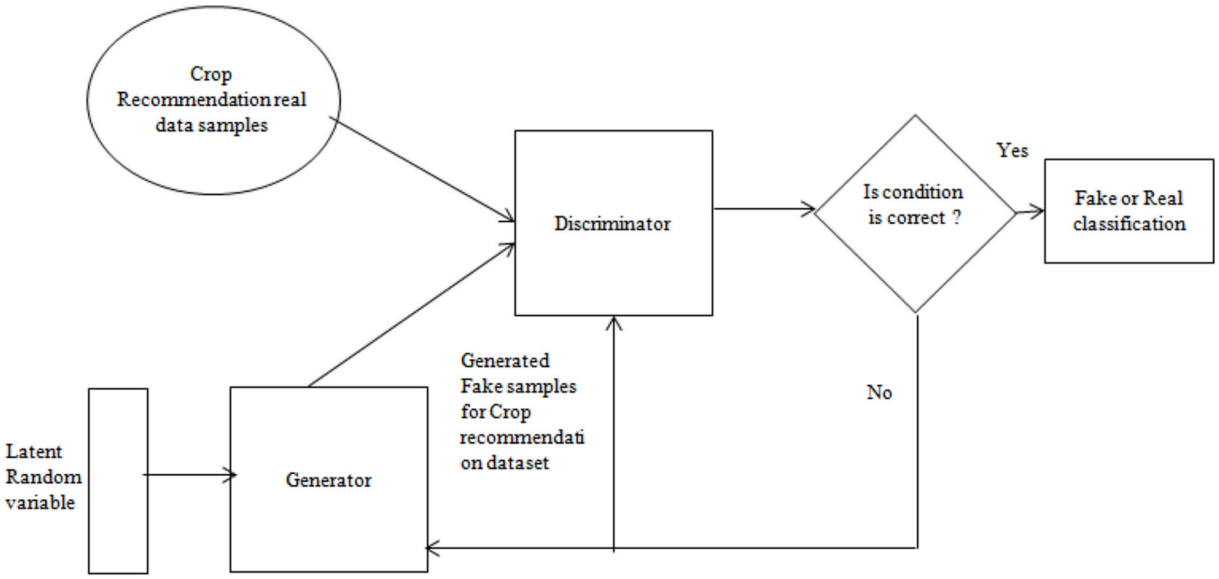

Fig. 1. General GANs architecture.

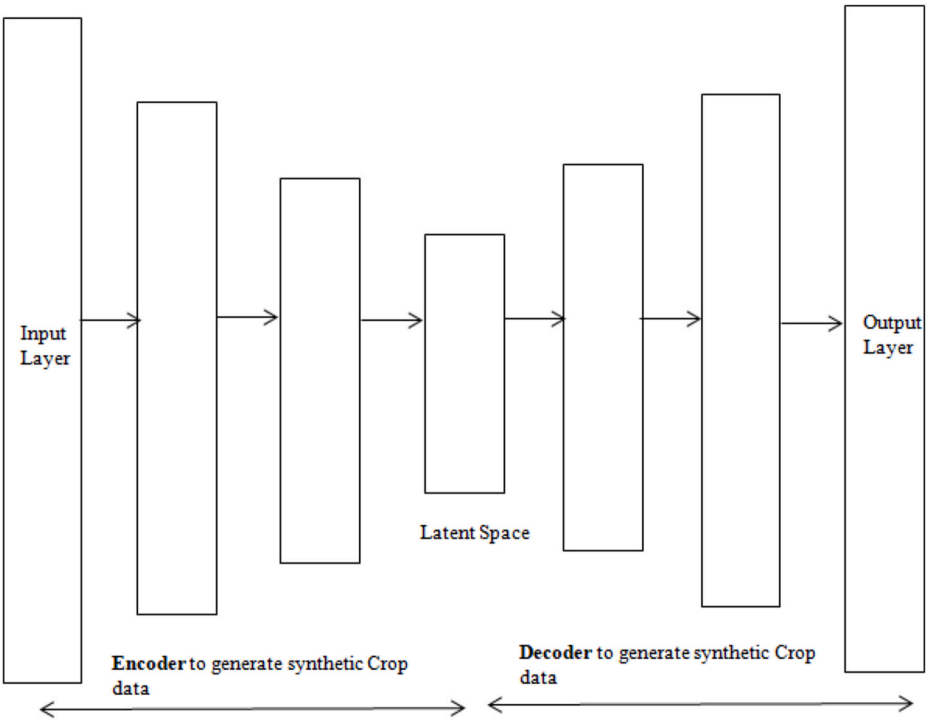

Fig. 2. VAE architecture.

3.4.2. Generative Adversarial Networks (GANs)

- Algorithm Complexity:

Generative Adversarial Networks (GANs) consist of two neural networks: a generator and a discriminator. The generator creates synthetic data, while the discriminator evaluates the authenticity of both real and synthetic data. The complexity of GANs is influenced by the architecture of both the generator and discriminator networks.

The computational complexity of GANs is O(g*n*d + d*n), where g represents the complexity of the generator network, n is the number of data points, and d is the dimensionality of the input data. The complexity can increase significantly with more complex network architectures or when training conditional GANs that require additional input parameters.

- Scalability:

GANs are known for their ability to generate high-quality synthetic data, but they can be computationally intensive, especially when dealing with high-resolution images or large datasets. GANs can also be trained using mini-batches, which makes them scalable to larger datasets. However, GANs are more challenging to train than VAEs due to issues like mode collapse and training instability. These issues can become more pronounced as the scale of the data increases.

To improve scalability, researchers have developed techniques such

Table 4
Comparison of GANs and VAEs.

Attribute	Generative Adversarial Networks (GANs)	Variational Autoencoders (VAEs)
Core Functionality	Generate data samples from random noise	Generate data samples with probabilistic encoding-decoding
Training Mechanism	Adversarial training between generator and discriminator networks	Encoder-decoder architecture with probabilistic encoding
Data Types	Often used for image, video, and audio generation	Widely used for image and data generation
Use Cases	Image synthesis, style transfer, super-resolution, and more	Image generation, data denoising, anomaly detection
Latent Space	Learns a continuous latent space for noise input	Learns a continuous probabilistic latent space
Interpretability	GANs are less interpretable due to their adversarial nature	VAEs offer some interpretability through the learned latent space
Data Quality Improvement	Useful for improving data quality by generating synthetic data	Useful for denoising data and generating realistic samples
Robustness to Input Noise	GANs can be sensitive to input noise and may produce inconsistent results	VAEs are generally more robust to input noise
Scalability	Scalable to high-resolution images but can be computationally intensive	Scalable for various data types with moderate computational requirements
Ethical Considerations	It can be used to create deep fakes and other potentially harmful content	It is less likely to be used for harmful content but still poses ethical concerns
Explainability	Often lacks explainability, making it challenging to understand model decisions.	Offers some explainability through the probabilistic latent space
Resource Requirements	Requires substantial computational resources, especially for high-quality image generation	Moderate resource requirements compared to GANs
Real-time Applications	It can be challenging to deploy real-time applications due to computational demands.	Suitable for real-time applications like anomaly detection
Notable Examples in Research	DCGAN, StyleGAN, BigGAN	VAE, β-VAE, Conditional VAE

as progressive growth of GANs, where the model starts with low-resolution images and gradually increases the resolution as training progresses. This approach can make the training process more stable and scalable to larger image datasets.

3.4.3. Practical implications for smart farming

In the context of smart farming, the scalability and complexity of VAEs and GANs are critical considerations. Smart farming datasets can be large and complex, with high-dimensional data from various sources, including satellite imagery, sensor data, and environmental measurements.

For practical applications, it is essential to balance the complexity of the models with the available computational resources. In many cases, simpler models that require less computational power may be preferred, especially in resource-constrained environments. However, for applications that demand high-quality synthetic data, such as precision agriculture and crop simulation, the use of more complex VAEs and GANs architectures may be justified.

When implementing these AI techniques in real-world agricultural operations, it is crucial to consider the trade-off between the quality of the generated synthetic data and the computational resources required. Additionally, the ability to scale these models as the dataset grows is vital for ensuring that the AI systems remain effective and efficient over time.

3.5. Simulation validation

Simulation validation is a critical step in ensuring that the models and methodologies proposed in a study are reliable and can be effectively applied in real-world scenarios. In the context of using Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) for generating synthetic crop recommendation data, simulation validation involves several key requirements and considerations.

3.5.1. Requirements for correct simulation

- Representative Data: The simulation must use data that accurately reflects the real-world conditions of agricultural settings. This includes a wide range of soil types, weather patterns, and crop types. The data should be sourced from reliable agricultural databases or collected through precise measurements in the field.
- Model Calibration: The parameters of the VAEs and GANs must be calibrated to ensure that the synthetic data generated is statistically similar to real-world data. This involves fine-tuning hyperparameters such as the number of layers, the size of the latent space, and the learning rate.
- Robustness Checks: The simulation should include robustness checks to evaluate the sensitivity of the models to changes in input data. This can be done by varying the input data and observing the changes in the output to ensure that the models are stable and provide consistent results.
- Comparative Analysis: The quality of the synthetic data should be compared with real data using various statistical measures such as correlation coefficients, mean squared error, and distributional similarity. This comparison helps in assessing the fidelity of the synthetic data.
- Scenario Testing: The simulation should test various agricultural scenarios, including extreme weather events, pest outbreaks, and varying market demands, to ensure that the synthetic data can capture the complexity of different farming conditions.
- Feedback Loops: Incorporating feedback from domain experts, such as agronomists and farmers, can help in validating the simulation. Their insights can guide the refinement of the models to better align with practical agricultural knowledge.
- Ethical and Legal Considerations: The simulation must adhere to ethical guidelines and legal regulations, especially when dealing with data that may have privacy implications or when the synthetic data is used for decision-making that affects livelihoods.
- Computational Feasibility: The simulation should be computationally feasible, with the ability to run on available hardware without requiring excessive computational resources that may not be accessible to all potential users.
- Scalability: The models should be scalable to handle larger datasets and more complex scenarios without a significant loss in performance or accuracy.
- Interpretability: The simulation results should be interpretable, providing clear explanations for the recommendations made by the crop recommendation systems. This is crucial for gaining the trust of end-users.

4. Pseudo code of VAEs and GANs to generate synthetic data for crop recommendation

The actual code for the VAEs synthetic data generation uploads at https://github.com.

/Yaganteeswarudu940/synthetc_crop_recommendation.

The summary of code mentioned in GitHub as follows. The crop recommendation dataset is loaded by using Python packages To handle

categorical features like 'N,' 'P,' 'K,' 'temperature,' 'humidity,' 'ph,' 'rainfall,' and 'label,' LabelEncoder is applied individually to each feature. This encodes categorical values into numeric representations, making them suitable for machine learning. The data is scaled using MinMaxScaler, which transforms feature values to a common range between 0 and 1. Scaling ensures that all features contribute equally to the model and aids convergence during training. A latent dimension (latent_dim) is defined, determining the dimensionality of the latent space in the VAEs. For crop recommendation, the latent dimension refers to the number of features in the dataset. The encoder network is created with an input layer (encoder_input) followed by a hidden layer (encoder_hidden) with 64 neurons and a ReLU activation function. Two additional layers, z_mean and z_log_var, represent the mean and log variance of the latent space. Using the reparameterization trick, a custom sampling function is defined to sample data points from the latent space. This step is crucial for generating diverse and meaningful synthetic data. The decoder network comprises an input layer for the latent space, a hidden layer (decoder_hidden) with 64 neurons and ReLU activation, and an output layer (decoder_output) with a sigmoid activation function. The decoder aims to reconstruct the input data from the latent space. The VAEs model is constructed by specifying the input and output layers, connecting the encoder and decoder networks. The VAEs model is compiled using the Adam optimizer and the custom loss function for training. The VAEs model is trained using the training data for a specified number of epochs and batch size. The validation data is used to monitor training progress. Synthetic data is generated by randomly sampling from the VAEs latent space. The number of samples matches the size of the original dataset. The synthetic data is inverse-transformed to its original scale using the MinMaxScaler, ensuring it aligns with the original data distribution. A data frame named 'synthetic_data' is created to organize and store the generated synthetic data, preserving the same feature structure as the original dataset. The encoded categorical features in the synthetic data are inverse-transformed back to their original labels using the respective LabelEncoders. The synthetic data is now ready for use in applications such as crop recommendation.

The actual code for the GANs synthetic data generation is uploaded at https://github. com/Yaganteeswarudu940/synthetc_crop_recommendation.

The algorithm begins by importing essential Python libraries for data manipulation, visualization, and deep learning. It loads a dataset and pre-processes it, renaming columns and encoding labels. The feature matrix and target variable are prepared, followed by handling missing values through K-nearest neighbors imputation and standardizing the data. Next, GANs class is defined with key methods: '_noise()' generates random noise in a latent space, '_generator()' defines the generator model, '_discriminator()' defines the discriminator model, and '_GAN()' combines them into a GANs model. The 'train()' method iteratively trains the GANs, with the discriminator distinguishing between real and synthetic data and the generator creating synthetic data. The GAN is instantiated, and the generator, discriminator, and GAN models are created and trained. Synthetic data is generated, and its quality is evaluated by comparing the original and synthetic data's correlation matrices and scatter plots.

5. Evaluation of generated crop recommendation synthetic data

5.1. Data set details

In essence, the success of this study in improving crop recommendation systems hinges on the premise that the synthetic data closely mirrors real-world crop data. The use of VAEs and GANs allows for the creation of data that should ideally represent the genuine conditions faced by farmers. However, if the training data used to teach these models is flawed, it could lead to the production of synthetic data that perpetuates inaccuracies and biases, ultimately undermining the goal of providing reliable recommendations to farmers. Therefore, the quality and accuracy of the training data are crucial considerations in the context of this research, and efforts should be made to ensure that the training data is as representative and unbiased as possible to maximize the effectiveness of the synthetic data in improving crop recommendation models.

The sample dataset used can be found at https://github.com/Yaganteeswarudu940/synthet c_crop_recommendation.

The research relies on real-world agricultural data, which includes information on various crops, soil conditions, and weather patterns. The data used in this research is collected from historical and current agricultural records around 10 years of data, in GitHub provided sample records, but actual records will be provided based on user request.

The amount of data used for training the VAEs, GANs is significant and depends on the complexity of the data and the model architecture. This research started with a thousand data points, and later, millions of data points are used to ensure the models can effectively learn and generate realistic crop data. To evaluate the quality and effectiveness of the generated synthetic data and recommendation models, a portion of the data (typically 20–30% of the total dataset) is set aside for validation and testing. This validation and testing data is crucial for assessing the model's performance and ensuring that the synthetic data aligns well with real-world scenarios. The research adopts a rigorous data-driven approach, combining advanced AI techniques with extensive real-world data to create high-quality synthetic crop recommendation data.

Currently, the following parameters are used for synthetic data generation, but in the future, there is a plan to use more soil parameters, weather parameters, and more. Nitrogen(N) is a macronutrient essential for plant growth. It is majorly responsible for the leaf development and the green color (due to chlorophyll, which has nitrogen). Phosphorus (P) is another macronutrient needed by plants. It plays a key role in the metabolic processes of plants, like photosynthesis, energy transfer and storage, cell division, and cell enlargement. Potassium (K) is another essential macronutrient for plants. It contributes to many vital functions in plants, including regulation of water uptake and loss, protein and starch synthesis, and the activation of enzymes. Temperature refers to the atmospheric temperature. It is crucial as it affects the rate of photosynthesis, germination, flowering, and fruiting in plants. Humidity is the amount of water vapor in the atmosphere. It affects many plant processes, including transpiration, photosynthesis, respiration, and plant pathogen interactions. pH(Potential of Hydrogen) This variable refers to the acidity or alkalinity of the soil. It has a direct effect on the nutrient availability in the soil, which the plants can absorb. Rainfall indicates the amount of precipitation in the form of rain. It is essential for plant growth as it provides a necessary water source. The amount of rainfall can largely affect the health and growth rate of crops; too little or too much rain can be harmful. It also helps in the dissolution and transportation of nutrients from the soil to the parts of the plant. Therefore, rainfall is considered while recommending crops because different crops require different amounts of water for their growth.

5.2. Evaluation approach

A heatmap is a graphical representation of data representing values as colors. In the context of a crop recommendation dataset, a heatmap can be used to visualize various aspects of the data and help make informed decisions about which crops are suitable for a particular region or set of conditions. One common use of a heatmap in crop recommendation is to visualize the correlation between different features or variables in the dataset. Each feature represents some aspect of the environment, such as temperature, rainfall, soil pH, humidity, etc. By creating a heatmap of the correlation matrix, you can quickly identify which features have a strong positive or negative correlation with each other.

Fig. 3(a) represents a heatmap for original data, Fig. 3(b) represents a heatmap for synthetic data generated for crop recommendation using

GANs, and Fig. 3(c) represents a VAEs synthetic data heatmap. In Fig. 3 (a), the correlation between feature 0 (nitrogen) and feature 1 (phosphorus) is −0.23, in Fig. 3(b), GANs synthetic data feature 0 (nitrogen) and feature 1 (phosphorus) correlation is −0.03, in Fig. 3(c) first and second feature correlation is −0.62. This compares all features with synthetic data, and after observing results, there is hyperparameter tuning required to make data to resemble actual data. So, in order to create similar synthetic data, GANs or VAEs models can be tuned with the number of epochs, for this iteration, 200 epochs are used, but the user can iterate with different epochs, 50 or 100 or different activation functions or different layers, and so on and compare heatmaps, which more approximate to original data can be considered for final synthetic data generation.

A scatter plot is a graphical representation used in data visualization to display the relationship between two continuous variables or dimensions in a dataset. In the context of a crop recommendation dataset, a scatter plot can be used to visualize how two specific variables or features relate to each other and can help in making informed decisions regarding crop selection. Scatter plots help in identifying outliers or unusual data points that don't conform to the general trend. Detecting outliers can be valuable in crop recommendations because they might represent unique conditions where a different crop choice is optimal. For instance, if there is a region with low rainfall but unusually high yields for a specific crop, that information can inform recommendations for similar conditions.

Fig. 4(a)represents a scatter plot for original data, Fig. 4(b) represents a scatter plot for synthetic data generated for crop recommendation using GANs, and Fig. 4(c) represents a VAEs synthetic data scatter plot. Fig. 4(b) and (c) data points are different from 4(a), so it is observed that the synthetic data generated is slightly different from the original data, so multiple iterations of epochs are required to generate an approximate crop recommendation dataset or hyperparameter tuning required to make synthetic data to resemble original data. Scatter plots are often used to display and analyze the correlation between two variables. The pattern of data points in the plot can illustrate whether there's a positive, negative, or no correlation between the variables.

A cumulative sum per feature plot, also known as a cumulative distribution plot, is a visualization technique used to understand the distribution of values within individual features (variables) in a dataset. This plot is particularly useful for crop recommendation datasets or any dataset where you want to explore the distribution of data within each feature. In a crop recommendation dataset, you typically have various features representing parameters like temperature, rainfall, soil type, and so on. Understanding the distribution of these features is crucial for making informed recommendations. The cumulative sum per feature plot helps you visualize how the data is distributed within each feature, which can be essential for understanding the range, central tendency, and spread of values. The plot also provides insights into how data is spread across the feature values. A widespread suggests variability in the environmental conditions, which may be important for understanding crop suitability.

Fig. 5(a) shows Nitrogen cumulative sums, 5(b) shows Cumulative sums of Phosphorous(P),5(c) shows Cumulative sums of Potassium(K),5 (d) shows Cumulative sums of temperature,5(e) shows Cumulative sums of humidity feature 5(f) shows Cumulative sums of Potential of Hydrogen(pH),5(g) shows Cumulative sums of rainfall,5(h) shows Cumulative sums of outcome label. All the figure shows the ability to generate synthetic data almost in a similar manner to the original data, but if need more approximate data requires running a model for different epochs.

A "Distribution per Feature" plot, also known as a feature distribution plot or histogram, is a valuable visualization tool for understanding the distribution of values within individual features (variables) in a dataset. Crop recommendation datasets typically include various environmental and agricultural factors as features (e.g., temperature, rainfall, soil pH). The distributions of these features can significantly impact crop suitability. Distribution per feature plots help you visualize how the data is distributed within each feature, allowing you to gain insights into the range, central tendency, and variability of values. This type of plot is typically displayed as a histogram, with the x-axis representing feature values bins or intervals and the y-axis representing the frequency or count of data points falling into each bin. The width of the distribution in

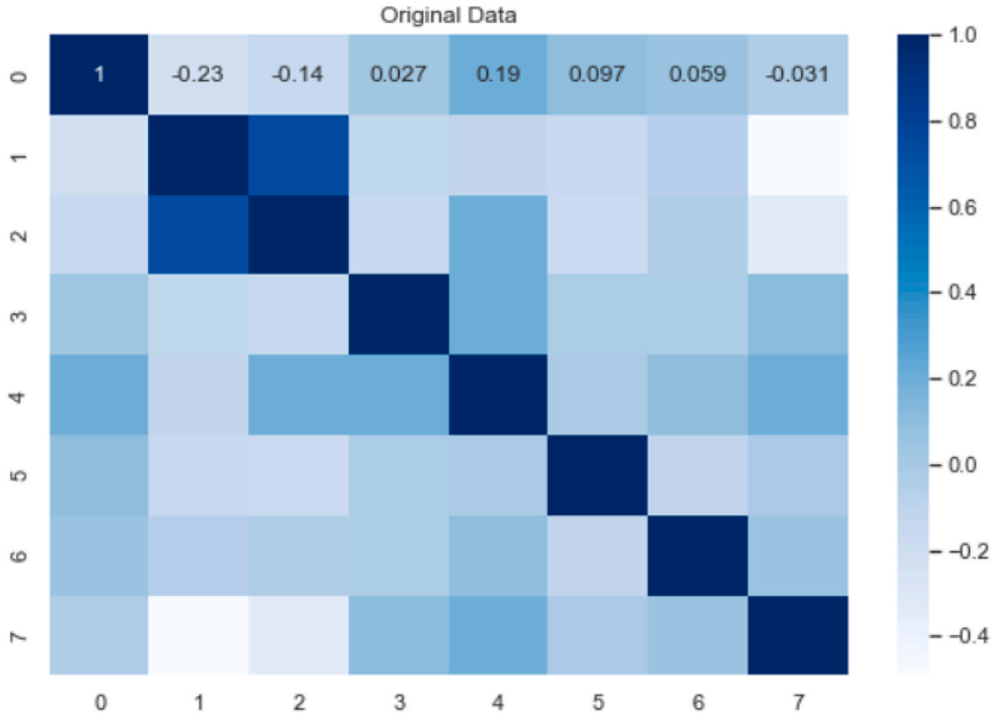

Fig. 3a. Original data heatmap.

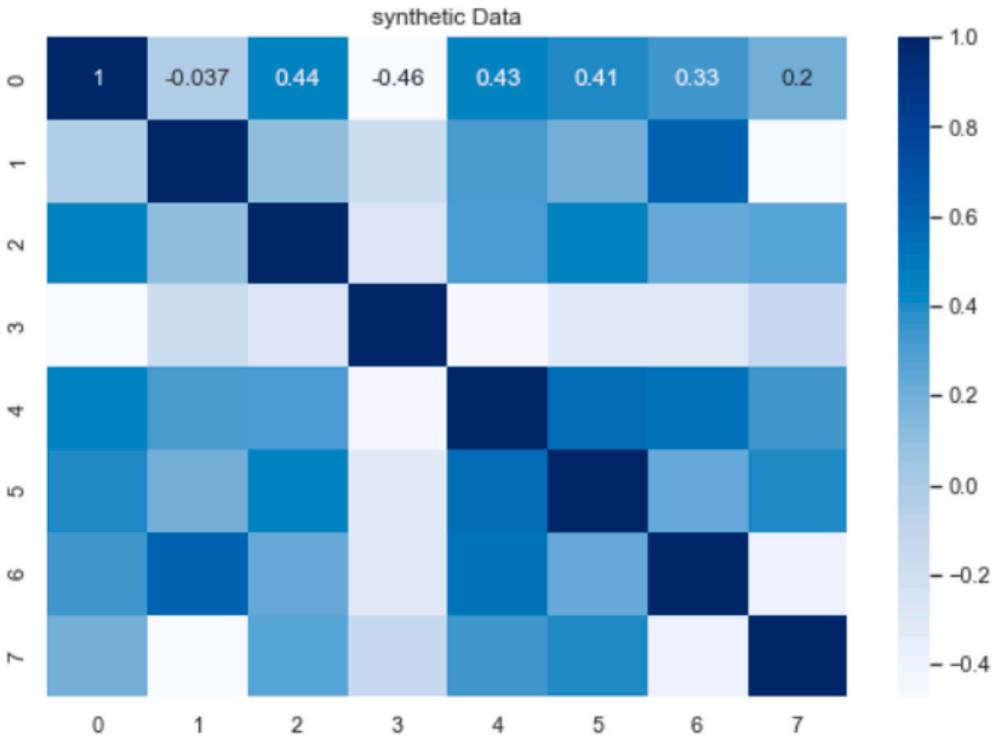

Fig. 3b. GANs Synthetic data heatmap.

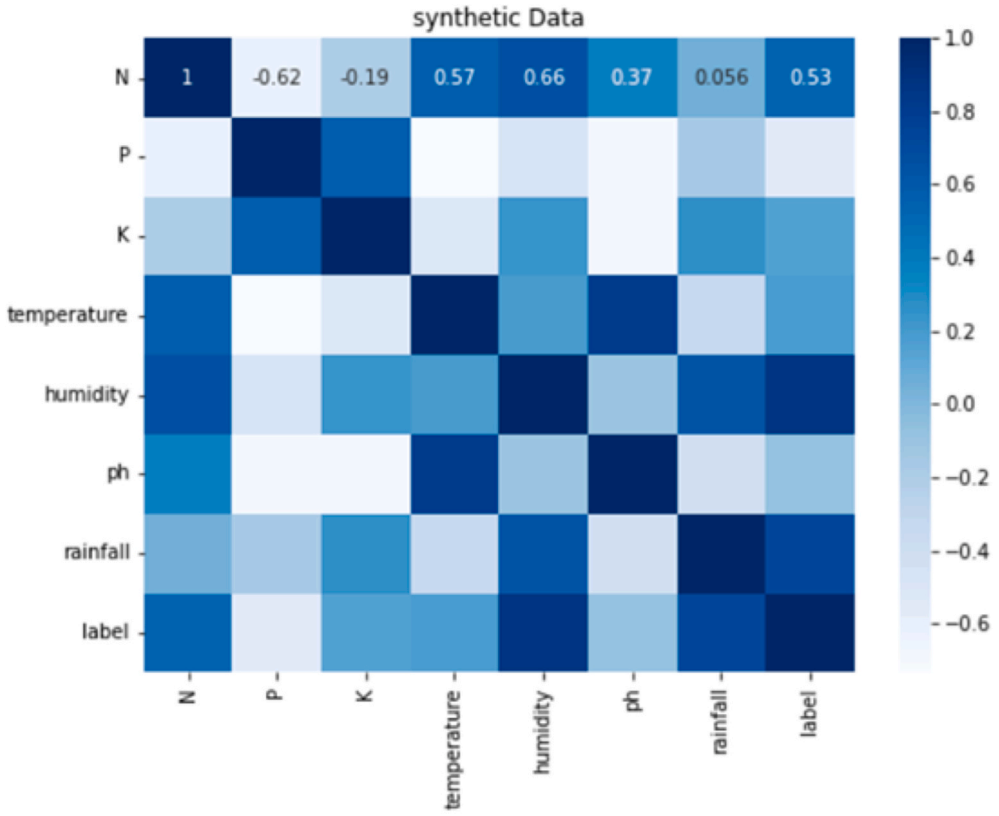

Fig. 3c. VAEs Synthetic data heatmap.

the histogram helps you understand the variability of feature values. A wider distribution indicates greater variability in environmental factors, which can be essential for crop suitability assessments. Fig. 6(a)–6(h) represent the Distribution of features Nitrogen, Potassium, and so on and also compare how original and synthetic data looks in each feature. Observing the above graphs by iterating the model for multiple epochs may get approximate data for the crop recommendation dataset.

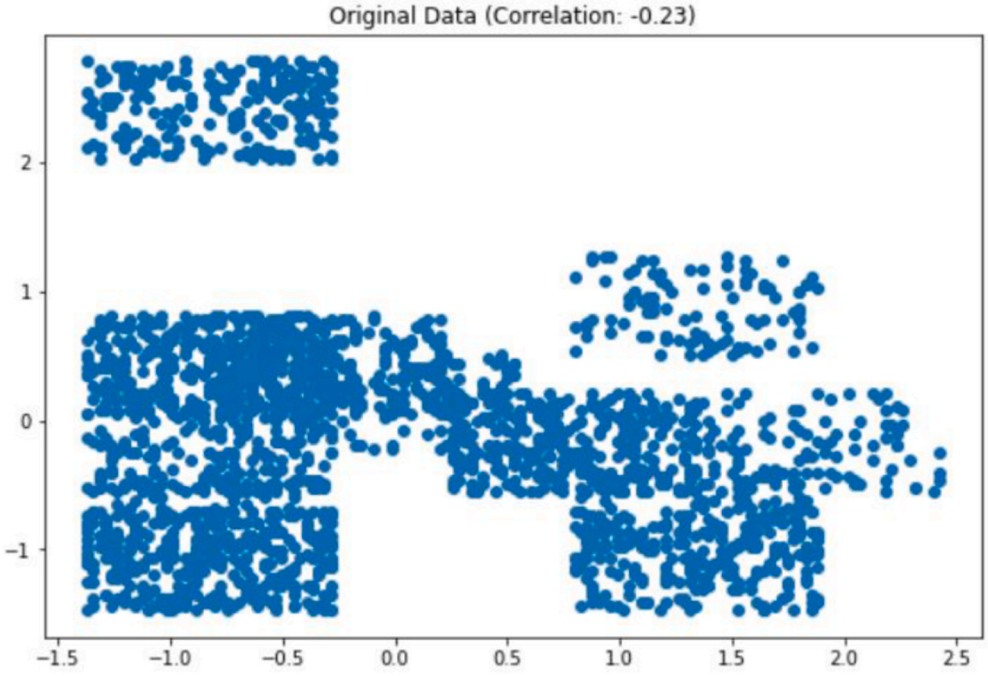

Fig. 4a. Original data scatter plot with correlation.

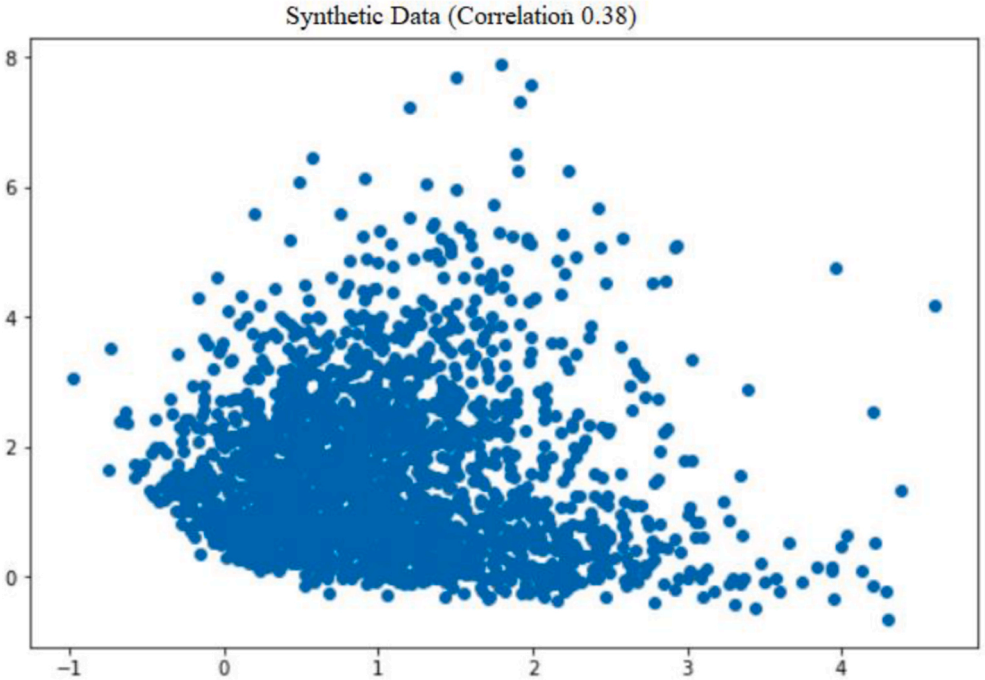

Fig. 4b. GANs Synthetic data scatter plot with correlation.

5.3. Limitations, assumptions, challenges, and hyperparameter tuning of the current study

Limitations of the current study include:

- Data Generalization: While the synthetic data generated using VAEs and GANs may closely resemble real-world crop data, there is a risk that the generated data may not fully capture the complexity and diversity of actual agricultural conditions. The ability of the synthetic data to generalize to various geographic regions, climates, and soil types should be carefully assessed.
- Model Assumptions: The effectiveness of VAEs and GANs in generating synthetic data depends on various assumptions and hyperparameters. If these assumptions do not align perfectly with the characteristics of the crop recommendation dataset, the generated data may exhibit biases or inaccuracies.
- Overfitting: There is a risk of overfitting when training generative models like VAEs and GANs. If the synthetic data generation models are not regularized properly, they may produce too specific data to

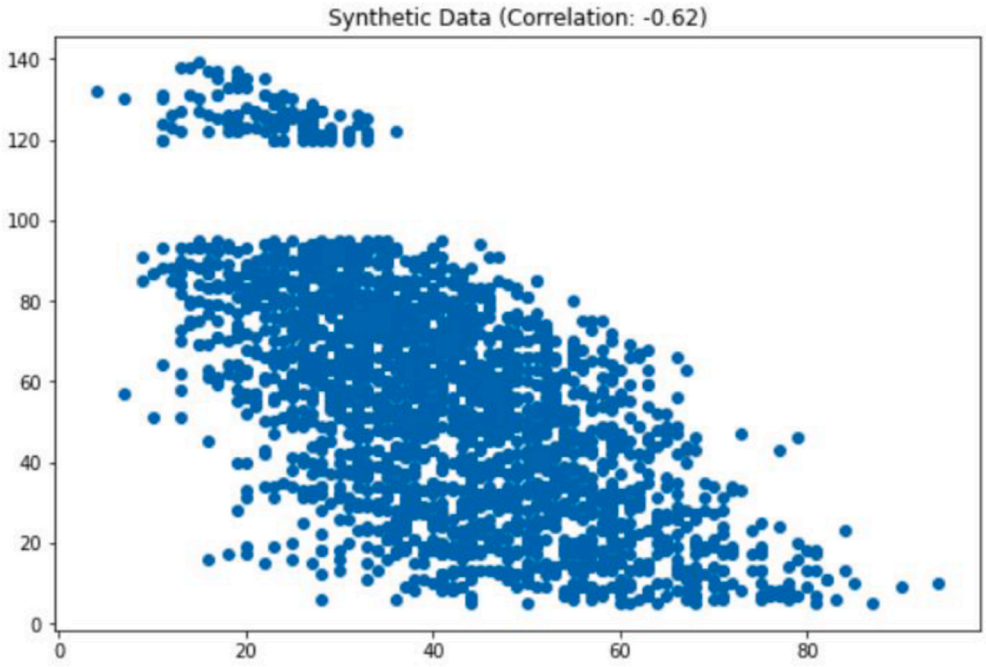

Fig. 4c. VAEs Synthetic data scatter plot with correlation.

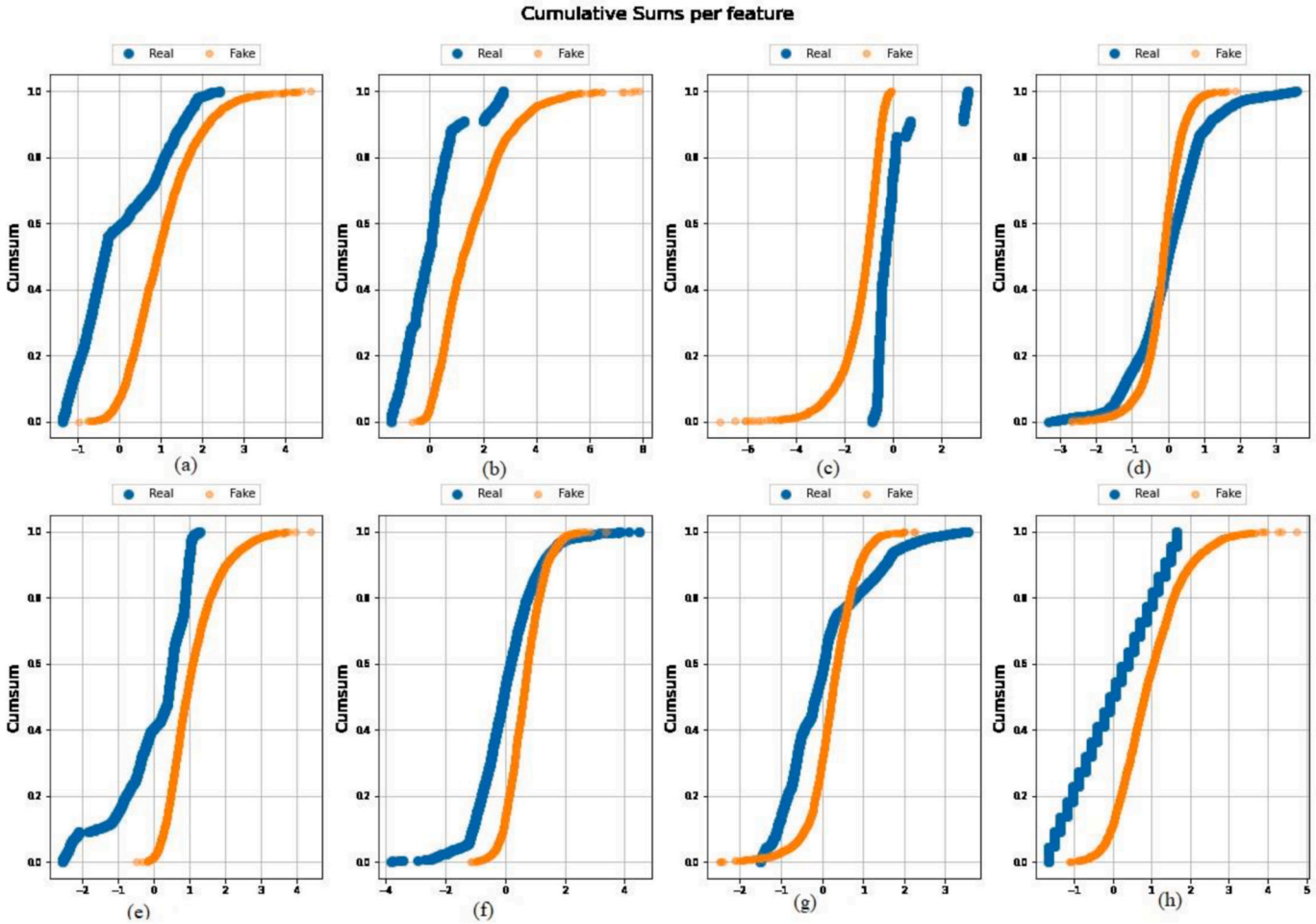

Fig. 5. **(a)** Cumulative sums per Nitrogen(N) feature **(b)** Cumulative sums per Phosphorous(P) feature **(c)** Cumulative sums per Potassium(K) feature **(d)** Cumulative sums per temperature feature **(e)** Cumulative sums per humidity feature **(f)** Cumulative sums per Potential of Hydrogen(pH) feature **(g)** Cumulative sums per rainfall feature **(h)** Cumulative sums per label.

the training dataset, limiting their usefulness for broader applications.

- Computational Resources: Training VAEs and GANs, especially on large agricultural datasets, can be computationally intensive and time-consuming.
- Real-world Validation: While the study aims to improve recommendation models using synthetic data, the ultimate test of its success lies in real-world validation. Any synthetic data generated should be validated after model design so that it can be used for further.
- Dependency on Input Data Quality: The quality of the synthetic data generated by VAEs and GANs heavily depends on the quality of the input data used for training. Suppose the input data is noisy, complete, and biased. In that case, it can positively impact the quality of the synthetic data, so before using any sample data for synthetic data generation, all outliers noisy data should be removed.

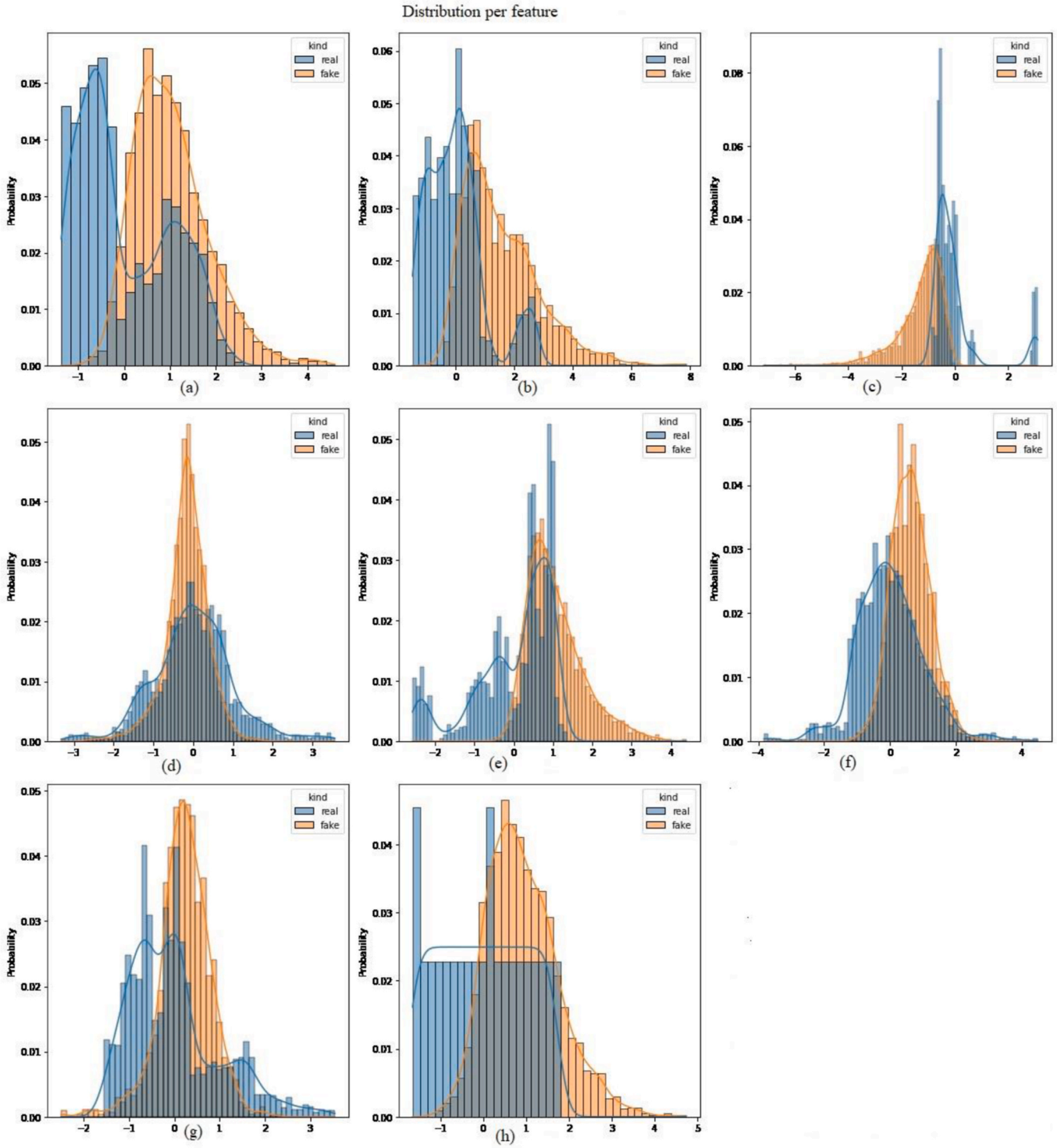

Fig. 6. **(a)** Distribution of Nitrogen(N) feature **(b)** Distribution of Phosphorous(P) feature **(c)** Distribution of Potassium(K) feature **(d)** Distribution of temperature feature **(e)** Distribution of humidity feature **(f)** Distribution of r Potential of Hydrogen(pH) feature **(g)** Distribution of rainfall feature **(h)** Distribution of label.

- Data Usage Restrictions: Some jurisdictions or organizations may have restrictions on the use of synthetic data for critical applications like crop recommendation. Legal and regulatory limitations on the use of synthetic data should be taken into account.

5.3.1. Challenges

One significant challenge lies in ensuring that the synthetic data accurately represents the diverse and complex conditions found in agricultural settings. While our models aim to closely mimic real data, there may still be areas where the synthetic data falls short, and this potential divergence needs to be carefully addressed. One significant challenge is the need for careful fine-tuning and parameter optimization of the VAEs and GANs to ensure the generated synthetic data is truly representative of real-world conditions. Moreover, the evaluation process itself can be challenging, as it requires the development of robust metrics and visualizations to assess the quality and utility of the synthetic data. These metrics should not only consider quantitative aspects but also qualitative factors, such as how well the synthetic data aligns with real-world agricultural scenarios.

Training deep learning models like VAEs and GANs can indeed be computationally intensive, and the feasibility of their use depends on several factors, including the scale of the dataset, available computing resources, and the specific problem being addressed. The use of VAEs and GANs for generating synthetic crop data can be resource-intensive, particularly if dealing with large and complex agricultural datasets. These models require significant computational power, which might not be readily available for all users or applications, especially in resource-constrained environments.

5.3.2. Assumptions

- Data Similarity: An assumption could be that the synthetic data generated by VAEs and GANs closely approximates the characteristics of real-world crop data. This assumption is fundamental to the success of the recommendation system.
- Relevance of Features: The assumption may be that the features used in the synthetic data, such as soil conditions, weather patterns, and market demand, are indeed relevant and critical for crop recommendation.
- Generalization: It could be assumed that the recommendation models built using synthetic data can generalize well to various agricultural regions and scenarios.

5.3.3. Hyperparameters

Fine-tuning hyperparameters is essential for optimizing the performance of Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). Unlike model parameters, which are learned from the data during training, hyperparameters are predetermined by the user to guide the learning process. They encompass aspects such as the design of the neural network and various factors involved in training. Vital hyperparameters to consider include the count of layers and neurons, the nature of activation functions, the learning rate, batch size, optimizer type, regularization and dropout rates, initialization method, loss function, epochs, learning rate schedule, momentum, batch normalization, weight constraints, and criteria for early stopping. The selection of these hyperparameters can vary based on the problem being solved, the dataset in use, and the structure of the network. Hyperparameters can be fine-tuned either manually or via techniques like grid search and random search, as well as automated procedures such as Bayesian optimization. This process generally involves trial and validation on a separate dataset to locate the optimal hyperparameters for a given task.

Among activation functions, the Rectified Linear Unit (ReLU) is widely used because of its efficiency in deep neural networks. It sets negative values to zero, enabling faster merging and alleviating the issue of vanishing gradients Premkumar et al., 2021; Houssein et al., 2022. Another activation function, the Sigmoid, is used in the final layer for binary classification tasks. It compresses input values into a range between 0 and 1 to denote probabilities. On the other hand, the Tanh activation function produces outputs between −1 and 1 and serves as a good alternative to the sigmoid function. Leaky ReLU solves the issue known as “dying ReLU”, which occurs when neurons stop learning during training by allowing a small slope for negative values. For multi-class classification issues, the Softmax function is typically used in the output layer. It transforms raw scores into probabilities for each class. The choice of activation function significantly impacts the network’s performance and stability during training, as it dictates how information moves within the network.

5.4. Comparison of state-of-art-work

The comparison with state-of-the-art work for the proposed use of Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) to generate synthetic crop recommendation (CR) data, we can draw parallels with current methodologies in synthetic data generation and their applications in various domains, as well as specifically within the agricultural context.

5.4.1. Comparison with current methodologies

- Variational Autoencoders (VAEs):
  i. State-of-the-art VAE methods have been extensively used for unsupervised learning and dimensionality reduction. These methods are adept at capturing the underlying probability distribution of the data. VAEs have seen success in domains like image processing, natural language processing, anomaly detection, etc.
  ii. The advantage of using VAEs for CR data generation is their ability to learn latent representations and interpolate between data points, potentially capturing complex agronomic relationships. However, VAE-generated data may sometimes be blurry or lack diversity compared to real data when applied to image generation tasks, which could translate to an over-generalization of crop conditions.
- Generative Adversarial Networks (GANs):
  i. GANs are at the forefront of synthetic data generation, most famous for their success in generating realistic images. They have also been used in data augmentation, domain adaptation, and privacy-preserving data generation.
  ii. GANs for CR could produce high-fidelity samples that respect the intricate relationships between soil properties, weather conditions, and crop performance. The pitfall, however, is that GANs require careful training to avoid issues like mode collapse, where the generated data lacks variety or instability during training.

Comparison with Agriculture-specific Applications:

- Synthetic Data for Crop Classification and Yield Prediction:
  i. Some studies in the agriculture domain have employed GANs for tasks like crop classification from satellite imagery and yield prediction with success.
  ii. Agricultural data often encompasses temporal and spatial complexities that may not always be captured by standard synthetic data generators. The use of VAEs and GANs specifically tuned for temporal and spatial consistency in CR data could represent a novel approach to solving this problem if current methodologies do not account for such factors.
- Data Augmentation in Agricultural Settings:
  i. Synthetic data aids in addressing data scarcity and imbalances in class representation, which is common in datasets concerning rare crops or extreme weather events. In these cases, augmentation

through synthetic data has proven beneficial for improving model performance.

ii. The proposed VAEs and GANs frameworks would need to demonstrate their effectiveness in generating balanced, diverse, and representative data for these less-representative scenarios while aligning with real-world distributions more convincingly than current models.

- Model Evaluation with Synthetic Data:
  i. Recent research has shown trends towards developing evaluation frameworks using synthetic data to assess the generalization and robustness of models, particularly when real labeled data is lacking.
  ii. A key challenge for the proposed study is to validate that the synthetic data produced by VAEs and GANs not only appears realistic in visualizations but also concretely enhances CR model performance properties like generalization, accuracy, and reliability.

In the comparison, the uniqueness of this study lies in its focus on crop recommendation systems, where the creation of synthetic data could tackle data limitations prevalent in the agricultural sector. While VAEs and GANs have been applied across various fields with significant achievements, their application to CR systems presents a niche area of exploration. This study could provide evidence of the efficacy of synthetic data to support agricultural decision-making processes through improved CR models. The empirical evaluation using visualizations and metrics to assess the synthetic data's quality will be crucial in establishing the generated data's credibility compared to state-of-the-art methods.

5.5. Real-world applications of synthetic data in smart farming

5.5.1. Case study 1: Precision irrigation

One practical implementation of smart farming is precision irrigation, which utilizes data-driven insights to optimize water usage for crops. By leveraging synthetic data generated through VAEs and GANs, farmers can simulate various irrigation scenarios under different weather patterns and soil conditions. This approach allows for the creation of a more efficient irrigation schedule that conserves water while ensuring crops receive the necessary hydration. For instance, a study in California used synthetic data to model the water needs of almond trees under different climatic conditions, leading to a 20% reduction in water usage without compromising yield.

5.5.2. Case study 2: Pest and disease management

Another application is in pest and disease management. Synthetic data can help in training models to predict outbreaks based on historical patterns of pest infestations and disease spread. For example, a vineyard in France employed a synthetic data-driven model to predict the risk of downy mildew, enabling the vineyard to apply fungicides more effectively and only when necessary, reducing chemical usage and environmental impact.

5.5.3. Case study 3: Crop yield prediction

Synthetic data is also valuable for crop yield prediction. By generating data that mimics various crop growth conditions, farmers can better anticipate yield outputs and plan accordingly. In Brazil, a soybean farm used synthetic data to predict yields under different planting densities and fertilizer applications, which helped optimize resource allocation and maximize profits.

5.5.4. Case study 4: Soil health monitoring

Soil health monitoring is crucial for sustainable farming practices. Synthetic data can be used to train models that predict soil nutrient levels, helping farmers apply the right amount of fertilizers. A cooperative in the Netherlands used a model trained on synthetic data to monitor soil health across different farms, leading to a more targeted and environmentally friendly fertilizer application strategy.

5.5.5. Case study 5: Crop planning and rotation

Effective crop planning and rotation are essential for maintaining soil fertility and preventing disease cycles. Synthetic data can assist in modeling the long-term effects of different crop rotations, taking into account market demand and climate change scenarios. An agricultural enterprise in the Midwest United States used synthetic data to optimize their crop rotation schedule, resulting in improved soil health and increased overall yields.

5.5.6. Case study 6: Market demand forecasting

Market demand forecasting is vital for ensuring profitability. Synthetic data can help predict future market trends, allowing farmers to plan their crop choices and quantities accordingly. A cooperative in Kenya used synthetic data to forecast the demand for various crops in local and international markets, enabling farmers to grow crops with higher market potential and secure better prices.

5.5.7. Case study 7: Greenhouse automation

Greenhouse automation is an area where synthetic data can play a significant role. By simulating different environmental conditions within a greenhouse, farmers can determine the optimal settings for temperature, humidity, and light to maximize plant growth. A greenhouse operator in Spain used synthetic data to automate their climate control systems, leading to a 15% increase in tomato production.

5.5.8. Case study 8: Supply chain optimization

Supply chain optimization is critical for reducing waste and ensuring fresh produce reaches consumers. Synthetic data can model various supply chain scenarios, from harvest to retail, to find the most efficient routes and storage conditions. A supply chain company in the United States used synthetic data to optimize the distribution of perishable goods, reducing spoilage rates by 10%.

5.5.9. Case study 9: Genetic crop improvement

Genetic crop improvement is another area where synthetic data can be beneficial. By simulating genetic variations and environmental interactions, researchers can predict which crop varieties will perform best under specific conditions. A research institute in India used synthetic data to accelerate the breeding of rice varieties that are more resistant to drought and pests.

5.5.10. Case study 10: Equipment maintenance and scheduling

Finally, equipment maintenance and scheduling can be enhanced with synthetic data. Predictive models can forecast when machinery will require maintenance, avoiding breakdowns during critical farming periods. A large farming operation in Australia used synthetic data to create a predictive maintenance schedule for their tractors and harvesters, reducing downtime by 25%.

5.6. The synthetic data broader implications for industries beyond agriculture

The methodologies discussed in this research, specifically the use of VAEs and GANs for generating synthetic data, have broader implications that extend well beyond the field of agriculture. These implications are significant across various industries where data scarcity, privacy concerns, and the need for robust data-driven decision-making are prevalent. Below are some of the key industries and applications where the methodologies could have a transformative impact:

- Healthcare and Biomedical Research:
  - Synthetic data generation can be used to create realistic patient records that maintain the privacy of individuals, enabling researchers to conduct studies without compromising sensitive information.
  - VAEs and GANs can help in drug discovery and personalized medicine by simulating patient responses to different treatments or drug combinations.
  - In medical imaging, these techniques can augment datasets for training machine learning models, improving diagnostic accuracy without exposing patient data.
- Finance and Economics:
  - Synthetic financial transaction data can be generated to test fraud detection systems, ensuring robustness without exposing real customer data.
  - Economic models can benefit from synthetic datasets that simulate various economic scenarios, aiding in policy-making and financial forecasting.
- Cyber security:
  - Generating synthetic network traffic data can help in training systems to detect anomalies and potential cyber threats.
  - Synthetic datasets can be used for penetration testing and security training, providing realistic scenarios for cybersecurity professionals to respond to.
- Automotive and Autonomous Vehicles:
  - Synthetic sensor data from VAEs and GANs can be used to train autonomous driving systems, enhancing their ability to navigate complex environments safely.
  - In crash simulations, synthetic data can help understand vehicle behavior under different conditions, leading to safer vehicle designs.
- Retail and E-commerce:
  - Synthetic customer data can be used to model shopping behavior and preferences, improving recommendation systems without using actual customer data.
  - Inventory management and demand forecasting can be optimized using synthetic datasets that model various market conditions.
- Manufacturing and Supply Chain:
  - Synthetic data can simulate production processes and supply chain disruptions, aiding in the development of more resilient manufacturing systems.
  - Predictive maintenance models can be trained on synthetic sensor data from machinery, preventing downtime and extending equipment life.
- Entertainment and Media:
  - In the creation of realistic digital content, such as video games and movies, GANs can generate lifelike textures and environments.
  - Music and audio production can benefit from synthetic data to create new sounds and compositions.
- Environmental Science:
  - Climate models can be enhanced with synthetic data to predict weather patterns and assess the impact of climate change on ecosystems.
  - Synthetic datasets can help in modeling the spread of pollutants and planning environmental remediation efforts.
- Education and Training:
  - Synthetic data can be used to create educational tools and simulations, providing students with realistic scenarios for learning.
  - In professional training, synthetic datasets can simulate workplace challenges, allowing for hands-on experience without real-world risks.

6. Conclusion

In this research, we addressed the challenge of limited labeled data for crop recommendation by harnessing the power of artificial intelligence techniques, specifically VAEs and GANs. Our primary goal was to generate synthetic crop data that closely mimics real-world agricultural conditions. We achieved this by utilizing VAEs to learn latent representations of input data and GANs to generate synthetic samples that are indistinguishable from real ones.

Our findings demonstrate the potential of these AI techniques in creating high-quality synthetic crop recommendation data. The synthetic data serves as a valuable resource for augmenting the crop recommendation dataset, enabling us to improve the performance of crop recommendation models. This research explored the effectiveness of VAEs and GANs in generating this data and presented the architecture and training processes of the proposed models.

In addition, we thoroughly evaluated the quality and usefulness of the generated synthetic data using various metrics and experiments, including visualizations like heatmaps, scatter plots, cumulative sum per feature plots, and distribution per feature plots. These evaluations provide insights into the utility and reliability of the synthetic data.

In future research, we can explore more advanced data augmentation techniques by combining VAEs, GANs, and other generative models to generate highly realistic synthetic crop data. These techniques may involve optimizing the model architectures, fine-tuning hyperparameters, and enhancing the diversity of the synthetic data to ensure it closely resembles real-world scenarios. This advanced data augmentation can lead to even more robust recommendation models.

Furthermore, the integration of user feedback and expert knowledge into the data generation process could be a promising avenue for enhancing the quality of synthetic data and recommendation models iteratively. By incorporating feedback from users, such as farmers and agricultural experts, we can adapt the generative models to better align with their specific needs and domain expertise. The overall effectiveness and usefulness of the recommendation systems can be enhanced by this iterative process, which can produce crop recommendations that are more precise and context-aware.

It's important to understand that although VAEs and GANs strive to create synthetic data that closely resembles real-world scenarios, the complexity of agriculture, with its myriad factors like soil conditions, weather patterns, and market dynamics, can be challenging to replicate accurately. This limitation has practical significance, especially when considering the application of these recommendation systems by farmers and stakeholders. Flawed or incomplete recommendations can lead to suboptimal decisions and potentially impact crop yields and profitability. It underscores the need for careful validation, constant monitoring, Parameter tuning, validation of synthetic data with original data with different graphs, and adaptation of the recommendation models to account for the inherent limitations of the synthetic data.

Future research can focus on developing more advanced architectures for VAEs and GANs that are specifically tailored to the complexities of agricultural data. This includes models that can handle temporal and spatial data, which are crucial for capturing the dynamics of crop growth and environmental changes. Smart farming generates various data types, including satellite imagery, sensor data, and textual information. Integrating these multimodal data sources using generative models could lead to more comprehensive synthetic datasets that better represent the multifaceted nature of agriculture. GANs, in particular, are known for training instability. Research into new training techniques and loss functions could make these models more stable and robust, leading to higher-quality synthetic data generation. As synthetic data becomes more prevalent, there will be a growing need to address ethical

and regulatory issues. Future research should explore frameworks for the responsible use of synthetic data, ensuring privacy, fairness, and compliance with agricultural regulations. There is a need for more transparent generative models that allow users to understand and trust the synthetic data generation process. Research into explainable AI (XAI) techniques could make these models more interpretable to farmers and other stakeholders. In agriculture, certain events like pest infestations or extreme weather conditions are rare but critical. Future research could focus on generating synthetic data for these rare events to improve the robustness of crop recommendation systems. Federated learning enables models to be trained across multiple decentralized devices holding local data samples. This approach could be used to generate synthetic data while preserving the privacy of individual farmers' data. Combining the strengths of VAEs and GANs with other machine learning techniques, such as reinforcement learning or transfer learning, could lead to more versatile and effective synthetic data generation methods. Generative models could be used to adapt data from one agricultural domain to another, making it possible to leverage data from different regions or crop types to improve crop recommendations. Future research should also focus on scaling synthetic data generation methods to support large-scale farming operations, which may involve vast datasets and require efficient processing. Developing generative models that can operate in real-time could enable dynamic crop recommendations that respond to immediate changes in environmental conditions or market demands. With the growing impact of climate change on agriculture, generative models could be used to simulate future scenarios and help farmers adapt their practices accordingly. Encouraging the creation of public agricultural datasets and open-source generative model frameworks could accelerate research and application in smart farming.

Statement of conflicting interest

The authors affirm that they have no known financial or interpersonal conflicts that may have influenced the research presented in this study.

Availability of data and code

The data used in the article and the code developed for the current study are available at https://github.com/Yaganteeswarudu940/synthetc_crop_recommendation.

CRediT authorship contribution statement

Yaganteeswarudu Akkem: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Writing – original draft, Writing – review & editing. **Saroj Kumar Biswas:** Software. **Aruna Varanasi:** Validation, Visualization.

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 paper.

Data availability

Data will be made available on request.

References

Agarwal, O., et al., 2021. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3554–3565.

Akkem, Y., Biswas, A.S.K., Varanasi, 2023a. Smart farming using artificial intelligence: a review. Eng. Appl. Artif. Intell. 120.

Akkem, Y., Biswas, S.K., Varanasi, A., Hassanien, A.E., Castillo, O., Anand, R., Jaiswal, A., 2023b. Smart farming monitoring using ML and MLOps. In: International Conference on Innovative Computing and Communications. ICICC 2023, vol. 703. Springer.

Andresini, G., et al., 2021. Autoencoder-based deep metric learning for network intrusion detection. Inf. Sci. 569, 706–727.

Asperti, A., 2019. About generative aspects of variational autoencoders. In: Machine Learning, Optimization, and Data Science—5th International Conference, LOD 2019, Siena, Italy, September 10–13, 2019, Proceedings, pp. 71–82.

Asperti, A., et al., 2021. A survey on variational autoencoders from a green AI perspective. SN Comput. Sci. 2, 301. https://doi.org/10.1007/s42979-021-00702-9.

Chen, S., Guo, W., 2023. Auto-encoders in deep learning—a review with new perspectives. Mathematics 11 (8), 1777. https://doi.org/10.3390/math11081777.

Chia, Y.K., et al., 2022. Relationprompt: leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. In: Findings of the Association for Computational Linguistics: ACL 2022, pp. 45–57.

Coutinho-Almeida, J., et al., 2021. GANs for tabular healthcare data generation: a review on utility and privacy. In: Soares, C., Torgo, L. (Eds.), Discovery Science. Springer International Publishing, Cham, Switzerland, pp. 282–291.

Dai, H., et al., 2023. Chataug: Leveraging Chatgpt for Text Data Augmentation arXiv preprint arXiv:2302.13007.

Diallo, B., et al., 2021. Deep embedding clustering based on contractive Autoencoder. Neurocomputing 433, 96–107.

Emam, K., et al., 2020. Chapter 1: introducing synthetic data generation. In: Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data. O'Reilly Media, Inc., Sebastopol, CA, USA, pp. 1–22.

Fan, Y., et al., 2020. Video anomaly detection and localization via Gaussian mixture fully convolutional variational autoencoder. Comput. Vis. Image Underst. 195, 102920.

Figueira, A., Vaz, B., 2022. Survey on synthetic data generation, evaluation methods and GANs. Mathematics 10 (15), 2733. https://doi.org/10.3390/math10152733.

Frid-Adar, M., et al., 2018. Synthetic data augmentation using GAN for improved liver lesion classification. In: Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 289–293. Washington, DC, USA, 4–7 April.

Goodfellow, I., et al., 2020. Generative adversarial networks. Commun. ACM 63, 139–144.

Hammouche, R., et al., 2022. Gabor filter bank with deep Autoencoder based face recognition system. Expert Syst. Appl. 197, 116743.

Han, Xu, et al., 2018. Fewrel: a large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. Proc. EMNLP.

Houssein, Essam H., Bahaa El-din, Helmy, Oliva, Diego, Jangir, Pradeep, Premkumar, M., Elngar, Ahmed A., Hassan, Shaban, 2022. An efficient multi-thresholding based COVID-19 CT images segmentation approach using an improved equilibrium optimizer. Biomed. Signal Process Control 73, 103401. https://doi.org/10.1016/j.bspc.2021.103401. ISSN 1746-8094.

Jain, N., et al., 2019. Agribot: agriculture-specific question answer system. [Online]. https://doi.org/10.35543/osf.io/3qp98.

Jangir, P., Buch, H., Mirjalili, S., et al., 2023. MOMPA: multi-objective marine predator algorithm for solving multi-objective optimization problems. Evol. Intel. 16, 169–195. https://doi.org/10.1007/s12065-021-00649-z.

Keskar, Nitish Shirish, et al., 2019. Ctrl: A conditional transformer language model for controllable generation. CoRR, arXiv:1909.05858.

Kim, K., Myung, H., 2018. Autoencoder-combined generative adversarial networks for synthetic image data generation and Detection of jellyfish swarm. IEEE Access 6, 54207–54214.

Li, X., et al., 2021. Weather gan: Multi-Domain Weather Translation Using Generative Adversarial Networks arXiv preprint arXiv:2103.05422.

Liu, Pengfei, et al., 2021. Pretrain, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. CoRR arXiv:2107.13586.

Liu, Y., et al., 2023. Summary of Chatgpt/gpt-4 Research and Perspective towards the Future of Large Language Models arXiv preprint arXiv:2304.01852.

Lu, G., et al., 2020. Multi-task learning using variational auto-encoder for sentiment classification. Pattern Recogn. Lett. 132, 115–122.

Meng, Yu, et al., 2022. Generating Training Data with Language Models: towards Zero-Shot Language Understanding. CoRR, 04538 arXiv:2202.

Pan, Z., et al., 2019. Recent progress on generative adversarial networks (GANs): a survey. IEEE Access 7, 36322–36333.

Park, N., et al., 2018. Data Synthesis Based on Generative Adversarial Networks. arXiv. arXiv:1806.03384.

Premkumar, Manoharan, Jangir, Pradeep, Sowmya, Ravichandran, 2021. MOGBO: a new Multiobjective Gradient-Based Optimizer for real-world structural optimization problems. Knowl. Base Syst. 218, 106856 https://doi.org/10.1016/j.knosys.2021.106856. ISSN 0950-7051.

Reynolds, Laria, McDonell, Kyle, 2021. Prompt programming for large language models: beyond the few-shot paradigm. Proc. CHI.

Rezayi, S., et al., 2022. Agribert: Knowledge-Infused Agricultural Language Models for Matching Food and Nutrition. IJCAI.

Saxena, D., Cao, J., 2021. Generative adversarial networks (GANs) challenges, solutions, and future directions. ACM Comput. Surv. 54, 1–42.

Siddani, B., et al., 2021. Machine learning for physics-informed generation of dispersed multiphase flow using generative adversarial networks. Theor. Comput. Fluid Dynam. 35, 807–830, 2021.

Xie, T., et al., 2021. Crystal Diffusion Variational Autoencoder for Periodic Material Generation arXiv 2021, arXiv:2110.06197.

Xu, L., et al., 2019. Modeling Tabular Data Using Conditional gan. arXiv. arXiv: 1907.00503.
Xu, L., Veeramachaneni, K., 2018. Synthesizing tabular data using generative adversarial networks arXiv 2018, arXiv:1811.11264.
Ye, Jiacheng, et al., 2022. Zerogen: Efficient Zero-Shot Learning via Dataset Generation. CoRR, 07922 arXiv:2202.
Zhang, G., et al., 2020. A survey of autoencoder-based recommender systems. Front. Comput. Sci. 14, 430–450.

Yaganteeswarudu Akkem Currently Ph.D. Scholar at NIT Silchar. Having 12+ years of experience in various industries. Reviewer for various reputed journals. Research interest includes machine learning, data science, AI and Artificial intelligence in agriculture. Published articles like smart farming using artificial intelligence : A review , smart farming monitoring using ML and MLOps, smart farming web application using machine learning approaches and many other articles. More than 20 papers are published in journals and international conferences.

Saroj Kumar Biswas completed a B. Tech Degree in Computer Science and Engineering from Jalpaiguri Govt. Engg. College, West Bengal, and M.Tech degree from National Institute of Technical Teachers' Training and Research, Kolkata. He completed his Ph.D. in computer science and engineering from NIT Silchar. He is currently an associate professor at NIT Silchar. His areas of interest are case-based reasoning, fuzzy logic, and machine learning algorithms.

Aruna Varanasi received a Ph.D. in CS E, JNTUH, and M. TECH (CSE) from Andhra University. Having a total of 24 years of experience (6 years Govt. service and 18 years of teaching). Areas of Research: Computer Networks, Cryptography, Image Cryptography, Information Security, and the Internet of Things. Published 1 book, 35 research papers in international journals, 5 patents filed, and 2 patents got published.