Privacy-Preserving Data Sharing in Agriculture: Enforcing Policy Rules for Secure and Confidential Data Synthesis

Anantaa Kotal
Dept. of C.S.E.E.
University of Maryland, Baltimore County
Baltimore, USA
anantak1@umbc.edu

Lavanya Elluri
Dept. of Computer Information Systems
Texas A&M University - Central Texas
Killeen, USA
elluri@tamuct.edu

Deepti Gupta
Dept. of Computer Information Systems
Texas A&M University - Central Texas
Killeen, USA
d.gupta@tamuct.edu

Varun Mandalapu
Dept. of Information Systems
University of Maryland, Baltimore County
Baltimore, USA
varunm1@umbc.edu

Anupam Joshi
Dept. of C.S.E.E.
University of Maryland, Baltimore County
Baltimore, USA
joshi@umbc.edu

arXiv:2311.15460v1 [cs.CR] 27 Nov 2023

Abstract—Big Data empowers the farming community with the information needed to optimize resource usage, increase productivity, and enhance the sustainability of agricultural practices. The use of Big Data in farming requires the collection and analysis of data from various sources such as sensors, satellites, and farmer surveys. While Big Data can provide the farming community with valuable insights and improve efficiency, there is significant concern regarding the security of this data as well as the privacy of the participants. Privacy regulations, such as the European Union's General Data Protection Regulation (GDPR), the EU Code of Conduct on agricultural data sharing by contractual agreement, and the proposed EU AI law, have been created to address the issue of data privacy and provide specific guidelines on when and how data can be shared between organizations. To make confidential agricultural data widely available for Big Data analysis without violating the privacy of the data subjects, we consider privacy-preserving methods of data sharing in agriculture. Synthetic data that retains the statistical properties of the original data but does not include actual individuals' information provides a suitable alternative to sharing sensitive datasets. Deep learning-based synthetic data generation has been proposed for privacy-preserving data sharing. However, there is a lack of compliance with documented data privacy policies in such privacy-preserving efforts. In this study, we propose a novel framework for enforcing privacy policy rules in privacy-preserving data generation algorithms. We explore several available agricultural codes of conduct, extract knowledge related to the privacy constraints in data, and use the extracted knowledge to define privacy bounds in a privacy-preserving generative model. We use our framework to generate synthetic agricultural data and present experimental results that demonstrate the utility of the synthetic dataset in downstream tasks. We also show that our framework can evade potential threats, such as re-identification and linkage issues, and secure data based on applicable regulatory policy rules.

Index Terms—Data Privacy, Privacy Policy, Privacy Attacks, Big data in Agriculture

I. INTRODUCTION

According to the Food and Agriculture Organization (FAO) [1], food production needs to be raised by 70% by 2050 to feed the projected population of 9.6 billion by 2050. To meet the growing needs of an expanding population, the farming community needs efficient ways to enhance agricultural productivity, optimize resource utilization, and implement sustainable farming practices. Technology plays an essential role in meeting these goals. Specifically, the use of Big Data in agriculture enables us to model large volumes of crowd-sourced data, sensor measurement data, and environmental information to understand emerging patterns in agriculture. This helps us stay ahead of modern challenges such as changing climate, farmland depletion, etc.

However, sharing data from multiple sources raises privacy concerns. Privacy regulations, such as the European Union's General Data Protection Regulation (GDPR) [2], have specific requirements on when and how such data can be shared. Even in the absence of specific regulations, individuals may have significant concerns about sharing their private information. We need privacy-preserving methods of sharing agricultural data that abide by policy regulations and encourage the farming community to participate in collective study.

The European Union's General Data Protection Regulation (GDPR) is one of the most comprehensive data protection regulations globally and has implications for agricultural data in EU member states. However, it does not specifically address data sharing in agriculture. The EU Code of Conduct on agricultural data sharing by contractual agreement [3] was created as an addendum to the GDPR. This code of conduct (Code) was designed to promote data-sharing leads in agriculture by setting transparent principles, clarifying responsibilities, and

creating trust among partners. The EU A.I. Act proposed in April 2021 by the European Commission proposed as the first EU regulatory framework for AI. It is the first regulation of its kind to define and classify AI used in different applications according to the risk they pose to users. AI models that pose higher risks require stricter regulation. In this study, the EU Code of Conduct on agricultural data sharing by contractual agreement [3], and the E.U. A.I. Act [4] have been identified as relevant privacy policies that can help us understand privacy concerns related to Big Data in agriculture.

Privacy in shared data has been discussed in prior work focusing on anonymization and encryption of data. However, these methods can be costly and not easily scalable. Privacy-preserving data generation refers to the process of creating new *synthetic* data that maintains privacy while retaining useful characteristics and statistical properties of the original data. Privacy-preserving data generation is a robust way of protecting sensitive data while still making them useful for wide-scale sharing. Generative Adversarial Networks (GANs) are one of the well-known models for generating synthetic samples that can have the same distributional characteristics as the original data. Synthetic data generated using privacy-preserving versions of GAN have been shown to replace real data for statistical and analytical purposes while protecting sensitive information [5]–[8]. While previous studies have successfully generated privacy preserving of synthetic data, they do not consider the requirements of privacy regulations. Thus, there is a gap between data privacy as defined by the privacy policies and the privacy constraints in the privacy-preserving methods.

To create a secure and privacy-preserving version of agricultural data that aligns with the regulations outlined in policy frameworks, we present an innovative framework. This framework is designed to generate data while strictly adhering to the guidelines established in agricultural privacy regulations. By incorporating the definition of privacy from these regulations into our model, we establish privacy constraints that specifically address the nuances of shared agricultural data.

The resulting synthetic dataset serves as a privacy-preserving alternative to confidential data, ensuring compliance with the principles outlined in agricultural codes of conduct. Our proposed framework effectively mitigates a range of threats, including privacy leakage, re-identification, side-channel, linkage, and attribute inference attacks. Through extensive experimentation, our model demonstrates resilience against various privacy threat models. Importantly, it achieves this while maintaining both statistical similarity to the original data and practical utility in downstream tasks.

The remainder of this paper is organized as follows. Section II establishes the background and motivation for our work in the context of potential threats against privacy-preserving framework, privacy policy regulations, and privacy-preserving data generation efforts. Section III presents our framework for policy enforcement in privacy-preserving data generation through rule extraction from privacy regulation and attribute regulation in generative models. Section IV presents the results
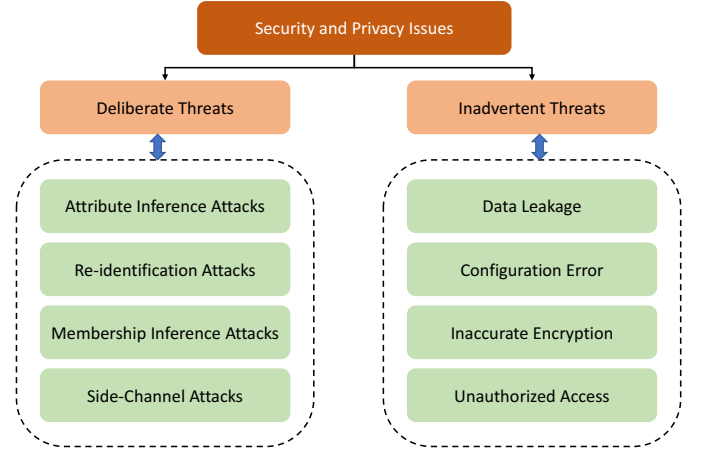

Figure 1: Overview of Potential Threats.

of our experimental work on policy-enforced data sharing Agricultural data. Section V presents the literature review on privacy regulations and methods to secure agriculture data from potential threats. The conclusion and future work are shown in Section VI.

II. BACKGROUND AND MOTIVATION

A. Privacy Threat Models in Agriculture

In the agricultural domain, privacy-preserving mechanisms are designed to protect individuals' sensitive information while allowing data to be analyzed and utilized for various purposes. However, these mechanisms can also be vulnerable to specific attacks that attempt to compromise the privacy they aim to preserve. Here are some potential threats against privacy-preserving measures, where threats are categorized as intentional and unintentional threats.

1) Deliberate Threats:

- Attribute inference attack: In attribute inference attacks, an attacker tries to reconstruct the original training data or extract sensitive information from the model's parameters. This type of attack is particularly concerning when the model is used to process private or personal data.
- Re-identification Attacks: Even if individual data points are anonymized, attackers could use auxiliary information from other sources to link seemingly anonymous data back to specific individuals, thereby breaching their privacy.
- Membership Inference Attacks: These attacks attempt to determine whether a specific data point was used during the training process. By analyzing the model's outputs, an attacker can infer whether a particular data point was part of the training dataset, potentially revealing sensitive information about the data owner.
- Side-Channel Attacks: These attacks focus on exploiting unintended information leakage during the execution of privacy-preserving algorithms. By analyzing execution

time, memory usage, or power consumption, attackers might gain insights into the data being processed.

2) Inadvertent Threats:

- Data Leakage: Confidentiality breach in agricultural data refers to a situation where sensitive or private information related to agriculture is unintentionally disclosed to unauthorized individuals or entities. This breach could involve the exposure of data such as farming practices, crop yields, land ownership details, financial information, and other sensitive information that farmers or agricultural organizations would prefer to keep confidential.

- Configuration Error: A configuration error as an unintentional error in agricultural data refers to a mistake or oversight made during the setup or arrangement of technology systems and software used in the agricultural context. These errors occur due to misconfigurations or incorrect settings that lead to unintended consequences or outcomes in the collection, storage, processing, or sharing of agricultural data.

  For instance, in precision agriculture where data-driven technologies are used to optimize farming practices, a configuration error could involve setting incorrect parameters for sensors, drones, or automated machinery. This might result in inaccurate data collection, leading to flawed decisions about irrigation, fertilization, or pest control.

- Inaccurate Encryption: Inaccurate encryption as an unintentional error in agricultural data refers to the incorrect implementation or utilization of encryption techniques meant to secure sensitive information related to agriculture. Encryption is a process of converting data into a secure, unreadable format to prevent unauthorized access or data breaches. However, when encryption is applied improperly or inaccurately, it can lead to unintended consequences and compromise the confidentiality of agricultural data.

  For instance, if encryption keys (the codes required to decrypt the data) are managed inadequately, there's a risk of unauthorized parties gaining access to the decrypted information. Additionally, using weak encryption algorithms or outdated encryption methods might render the data susceptible to decryption by attackers with sufficient computing power or knowledge.

  Inaccurate encryption practices in agriculture data could also involve failing to encrypt all necessary data fields or overlooking specific data sources, leaving certain parts of the information vulnerable to exposure. This can be particularly concerning when dealing with sensitive data such as crop yield projections, land ownership details, or financial records.

- Unauthorized Access: Unauthorized access as an unintentional error in agricultural data refers to situations where individuals or entities gain entry to sensitive agricultural information without proper authorization or permission. This error occurs due to vulnerabilities in data security measures, misconfigured access controls, or inadvertent lapses in safeguarding agricultural data.

  For instance, if an agricultural database containing information about crop yields, pricing strategies, or proprietary farming techniques lacks proper access controls, unauthorized individuals could gain access to this information. This could happen due to weak passwords, lack of encryption, or overlooking permissions that restrict data access to only authorized personnel.

  Unintentional unauthorized access might also occur if a legitimate user inadvertently shares login credentials or access links with unintended parties or if attackers exploit a software vulnerability to gain entry to sensitive agricultural databases.

B. Privacy Regulations in Agriculture

Code of Conduct describes contractual relations and guides the use of agricultural data, particularly on the rights of use and access of the data. Code was a collaborative step between massive institutions representing various enterprises producing animal fertilizers, seeds, feed, or farm machinery and agents of animal breeding organizations as well as farmer's cooperatives in the EU associated with the Council on Ethical & Judicial Affairs (CEJA) [9] and Copa-Cogeca (that concentrates on young farmers up to 40 years of age). In establishing the Code, it was highlighted by the parties that "the Code promotes the advantages of data sharing and allows agribusiness models, including agri-cooperatives and other agri-businesses, to shift into an age of digitally enhanced farming swiftly." [3]

The European Parliament has also launched a substantial amount of work in AI. Recently, "The AI ACT" [4] has been proposed to protect user's data from AI risks. The European Parliament has also launched a substantial amount of work in AI. Due to extreme technological transition in recent years and potential challenges, the EU desires balanced practices. The Union aims to preserve the EU's technological supervision and assurance that Europeans can benefit from new technologies designed and performed according to Union rules and regulations. Figure 2 shows the relation between the EU Code and the AI Act.

C. Synthetic Data in Privacy

Various approaches have been proposed for privacy-preserving data sharing. Some common approaches towards privacy use differential privacy [10], K-anonymity [11], L-diversity [12] and t-closeness [13]. Privacy-Preserving Generative Models use Generative models like Generative Adversarial Networks (GANs) to generate synthetic data that closely mimics the original data distribution. This synthetic data can be shared without revealing actual individual data thus providing privacy.

Generative adversarial Networks (GANs) is a class of deep learning-based generative model in AI. GANs are extremely accurate in synthetic data generation and translation, particularly for image and text data [14]–[17]. The principal architecture in a GAN framework involves a generative model
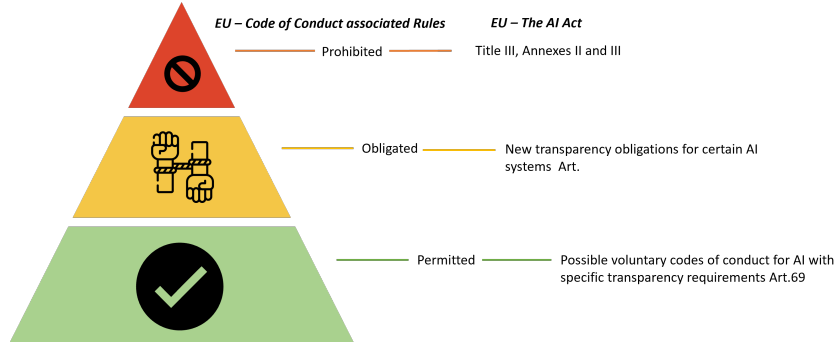
Figure 2: EU Code of Conduct Vs. AI Act

G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the original distribution rather than G. The training procedure for G is to maximize the probability of D making a mistake. As the training progresses, the generator gets better at generating new examples that plausibly come close to the samples from the original distribution. The idea behind GAN can be formulated as a two-player min-max game with value function $V(G, D)$:

$$\min_G \max_D V(G, D) = E_{x \sim p_{data}(x)}[log D(x)] \\ + E_{z \sim p_z(z)}[log(1 - D(G(z)))] \quad (1)$$

For agricultural data, the data is usually tabular i.e. a mix of discrete and continuous values. Additionally, the continuous values are not arbitrarily random and usually follow a specific distribution within a given range. To account for this, we need specialized versions of GAN that can accurately replicate system data that is collected over our digital twin. In an unconditioned generative model, there is no control over the modes of the data being generated. Conditional Generative Adversarial Nets (CGAN) [18] introduces the concept that by conditioning the model on additional information, it is possible to direct the data generation process. The objective function of the two-player minimax game is rewritten as:

$$\min_G \max_D V(G, D) = E_{x \sim p_{data}(x)}[log D(x|y)] \\ + E_{z \sim p_z(z)}[log(1 - D(G(z|y)))] \quad (2)$$

To ensure that the synthetic dataset is distributionally close to the original dataset, and provides privacy protection by the principle of t-closeness, Kotal et al. [8] propose the use of Earth Mover's distance (EMD). he The EMD of the distribution of features in the synthetic is calculated w.r.t. the original dataset. The sampling process continues to sample from the trained generator till the generated distribution is within a threshold distance of the original distribution. To address the challenges of tabular data, the model uses three key steps during generation: (1) Mode-specific normalization, (2) Conditional Generator, and (3) Training by sampling.

III. PRIVACY PRESERVING DATA GENERATION WITH POLICY ENFORCEMENT

The various models proposed for privacy-preserving data sharing do not take into regulations on Information Privacy. While they can provide privacy protection for data, there is still a gap between privacy as ensured by these models and privacy as required by the law. Bridging this gap requires translating legal privacy requirements into machine-enforceable values and mechanisms. This involves understanding the key principles of data protection regulations and incorporating them into designing and implementing privacy-preserving models. This work proposes a novel framework for Policy Enforcement in privacy-preserving data generation. We process the relevant privacy regulations and extract privacy rules from the human written document. The rules are converted into a machine-readable and enforceable format. The privacy rule sets are then used to guide the privacy-preserving data generation model. Figure 3 demonstrates the overall architecture of our proposed framework.

A. Rule Extraction from Privacy Regulations

We use a Deontic Logic Rule-based approach to extract relevant rules from Privacy regulations into a machine-enforceable format. First, we create a predefined list of modal verbs used to express obligation types in legal domains. Next, we tokenize each token in the set and use the Python library to get dependency analysis and POS (part of speech) tags. The algorithm searches for predefined triggers within a given sentence to extract its position (starting index), each mention of the trigger, and its dependency tags. Deontic modality [19] [20] is predominantly used in legal documents to describe vendors' prohibitions, obligations, and permissions. For instance, 'may', 'must', and 'must not', frame 'permission', 'obligation', and 'prohibition' respectively. Table I defines all the deontic types, and Table II has all the model verbs associated with each deontic type. Modality guides the linguistic ability to express alternative forms the world could be and is generally represented by modal auxiliaries such as must, can, shall, will, and may. Below are some examples from the EU Code of Conduct for agriculture.
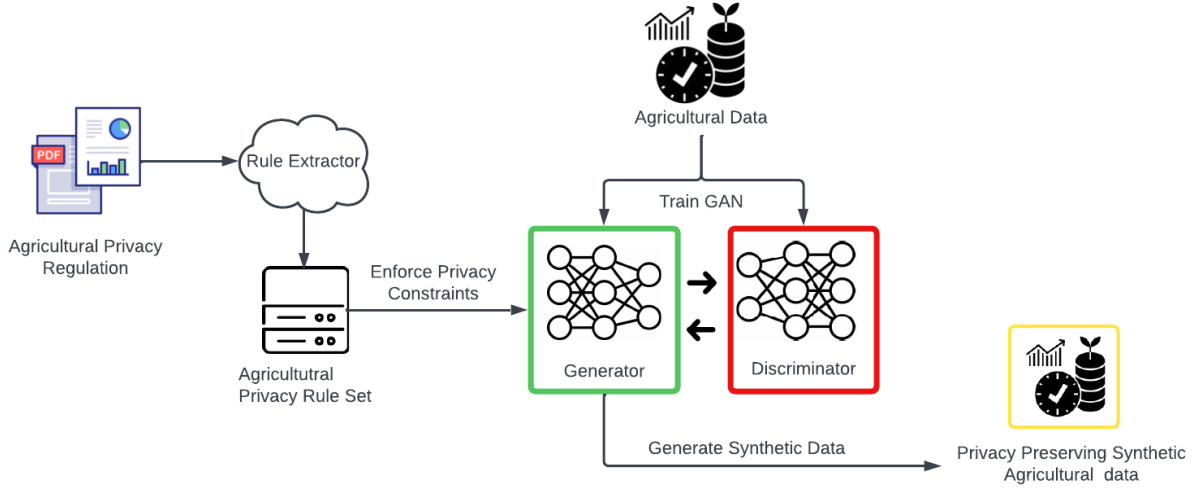
Figure 3: Framework for Privacy-Preserving Data Generation in Agriculture with Privacy Policy Enforcement

Permission Rule 1: The farmer can provide data to land owners, potato processors, the government, paying authorities, etc. [3]

Permission Rule 2: Unless otherwise agreed in the contract, the data originator can transmit this data to another data user. [3]

Obligation Rule 1: Contracts must not be amended without the prior consent of the data originator. [3]

Obligation Rule 2: Parties may not use, process, or share data without the consent of the data originator. [3]

Prohibition Rule 1: Data cannot be owned in the same way as physical assets. [3]

Prohibition Rule 2: Parties may not use, process, or share data without the consent of the data originator. [3]

Table I: Deontic Types Taxonomy

Deontic Modality	Definition
Permission	Vendor is authorized to do something.
Prohibition	Vendor is forbidden to do something.
Obligation	Vendor is mandated to do something.
Entitlement	Vendor has the right to do something.

B. Policy enforcement in Data Generation

The rules extracted from policy documents in Section III-A form a rule set. This rule set is used to guide the data generation process. We use the Privetab model [8] for privacy-preserving data generation. The principle of privacy used in the model is t-closeness [13] which stipulates that the distributional similarity between the original and synthetic data should be within a threshold to ensure privacy. The threshold value is not pre-determined and often determined based on the needs of an organization. To ensure that these privacy constraints meet the requirements of the privacy regulations, we use two key steps:

Table II: Modal verbs for Deontic Types

Type	Model Verbs
Obligation	shall be required, will be required, shall be obligated, shall, must, will, have to, should, ought to have, will be paid, shall be paid, agree, agrees, acknowledges, acknowledge, represents and warrants, shall be responsible for, will be responsible for
Prohibition	shall not, will not, must not, may not, cannot, shall have no right, can not, shall not be allowed, will not be allowed, shall not assist, shall be prohibited, will be prohibited, nor shall, not to be, neither lessor nore lessee may, in no event shall, nor will, will not allow, nor may
Permissions	shall be permitted, shall also be permitted, can, may, could, shall be allowed, will be allowed, is permitted, will allow, has the right, or at landlord's option, shall be permitted to

1) Determine Attribute Sensitivity: According to the EU Code of Conduct and "The AI Act", all the features in the dataset are associated with the relevant rule and risk level. We use the rule set extracted from the privacy policy to determine the risk level of each attribute in the original data. Attributes are categorized into three levels of sensitivity: **low**, **medium** and **high**. Highly sensitive attributes are extremely risky to share and thus require the highest privacy protection. Attributes with low sensitivity are less risky and thus can be shared with minimal risks.

2) Enforce privacy threshold based on Sensitivity category: When generating the privacy-preserving synthetic dataset, we have to ensure that the privacy threshold is satisfied. The value of the privacy threshold can be tuned. Based on the attribute sensitivity of the dataset determined from privacy regulations,

we modulate the privacy threshold for each attribute. The high-risk attributes have the strictest threshold for privacy. The privacy threshold for low-risk attributes is less strict. This ensures that the privacy requirements of the regulations are met in this process while maintaining the utility of the generated data for downstream tasks.

IV. EXPERIMENTAL RESULTS

Our framework for privacy-preserving data generation with policy enforcement can be used for sharing sensitive datasets. We demonstrate the use of our framework for Agricultural Big Data as a proof-of-concept. In this section, we use our framework to generate a privacy-preserving version of the ITM4Impact dataset [21]. The implementation of our policy-enforced privacy-preserving data generation framework is made available in a library[1]. We provide the experimental results to demonstrate that our framework provides a privacy-preserving alternative for sensitive datasets that is still useful in downstream tasks.

A. Dataset

In this experiment, we use the ITM4Impact dataset [21] collected by the ILRI institute to measure and determine the impact of infection and treatment method (ITM) on farmers. This dataset was anonymized for public release. We observe that some of the information shared in this dataset is privacy sensitive and poses a risk towards re-identification of participating farmers. In the ITM4Impact dataset, there are around 800 attributes. Approximately more than 140 groups segregate these attributes into multiple categories. According to the EU Code of Conduct and "The AI Act", all the features in the dataset are associated with the relevant rule and risk level. More than 500 attributes are related to the farmer PII data or the data generated in the farming environment and categorized with a "High" privacy level. More than 270 attributes, as per regulation, fall under "Medium" risk, and only around 30 features can be public and can be shared with the public. However, we see that many breaches have happened often, and people end up paying huge penalties for not securing the information by adhering to the regulation policies. As mentioned in the regulatory document, it is crucial that this data should be secured and cannot be shared unless the individual has permission.
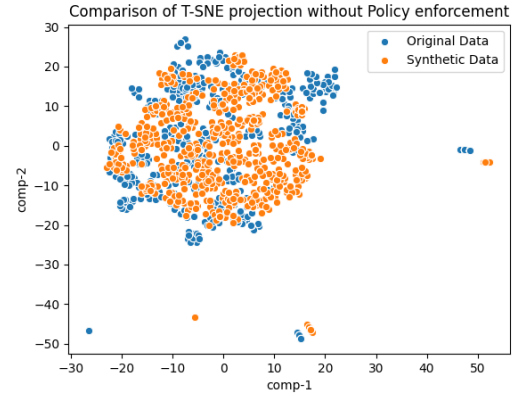
Below are some of the rules extracted from the regulation for applying them to the relevant attributes in the dataset:
The data originator can store data in a primary location, in a data platform, or cloud-based storage platforms. [3]
The datasets should only be kept for as long as is strictly necessary for the relevant analyses to be carried out. [3]
If the data is being used to make decisions about the data originator "as a natural person" the GDPR applies. For instance, the rights regarding data produced on the farm or during farming operations are granted to ("owned by") the farmer and may be used extensively by them. [3]

[1]https://github.com/Ebiquity/policy_enforced_data_generation
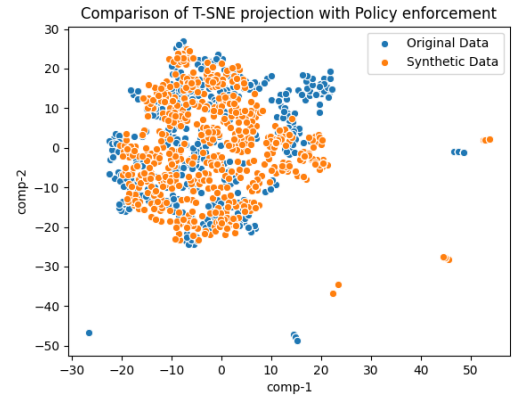
B. Fidelity Results

Evaluating the fidelity of synthetically generated data is a critical step in assessing the effectiveness of your privacy-preserving data generation framework. By measuring the realisticness of synthetic data, we can determine how closely it resembles real data and its potential for utility in downstream tasks. We provide two fidelity metrics for our model.

1) T-SNE Visualization: For utility in downstream tasks, it is important that the synthetic dataset provides the same data coverage as the original dataset. As observed in Figure 4 for privacy-preserving data generation, there is minimal loss in coverage from the original dataset from both with and without policy enforcement in the generation process.

(a) Comparison of T-SNE projection of Synthetic Data without Policy Enforcement

(b) Comparison of T-SNE projection of Synthetic Data with Policy Enforcement

Figure 4: Comparison of T-SNE projection of Original and Privacy-preserving Synthetic Data

2) Statistical Similarity: We provide an of the statistical resemblance between the original and synthetic datasets for attributes at the 3 different risk levels. In general, the metrics demonstrate a strong consistency between the synthetic datasets generated with our framework and the original dataset. Figure 5, 6 and 7 shows the cumulative distribution function (CDF) graphs comparing the original and synthetic data for
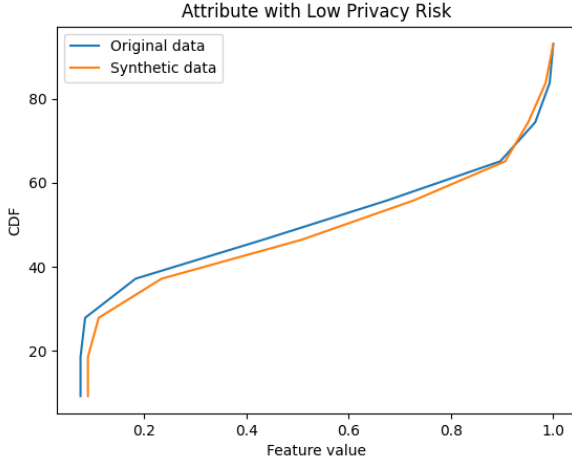
Figure 5: Comparison of CDF for Original vs Synthetic data in Low Risk Attribute
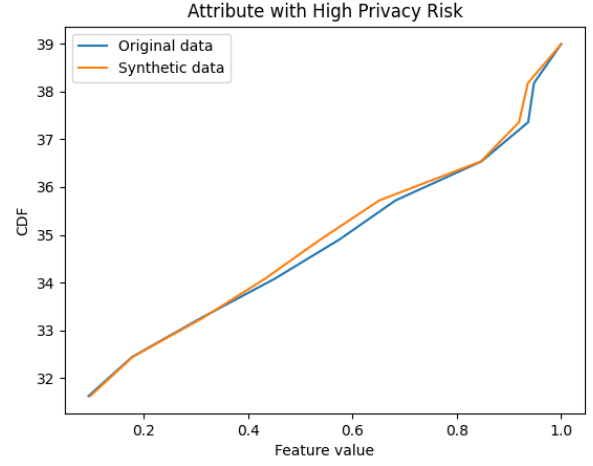

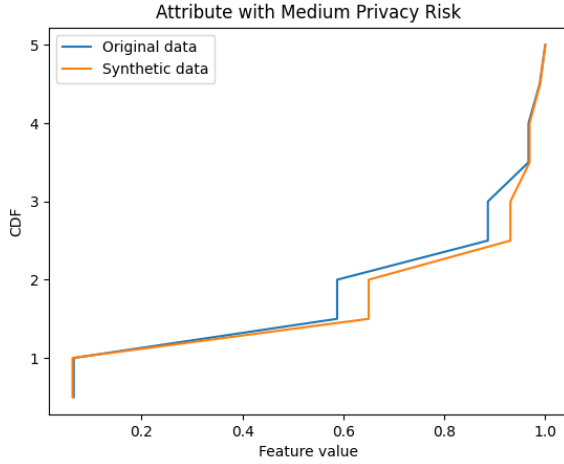Figure 7: Comparison of CDF for Original vs Synthetic data in High Risk Attribute

Figure 6: Comparison of CDF for Original vs Synthetic data in Medium Risk Attribute

three attributes at low, medium, and high privacy risk level respectively. The average KS statistic, which measures the maximum difference in CDFs between the original and synthetic data is 0.04 for all 3 attributes.
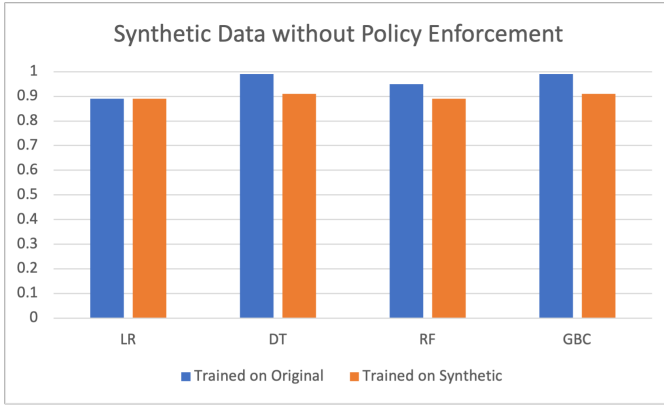
C. Utility Results

It is important that the resultant dataset of our framework is still useful in replacing the original dataset for downstream tasks. One of the most important use cases of big data is in the field of Machine learning (ML). We show here that the data generated from our framework can replace the original dataset for ML prediction tasks. We evaluated ML models trained on the original dataset against ML models trained on the generated data. Consistent performance across both models would signify that the synthetic data retains relevant information for the task. It is important to note here the test
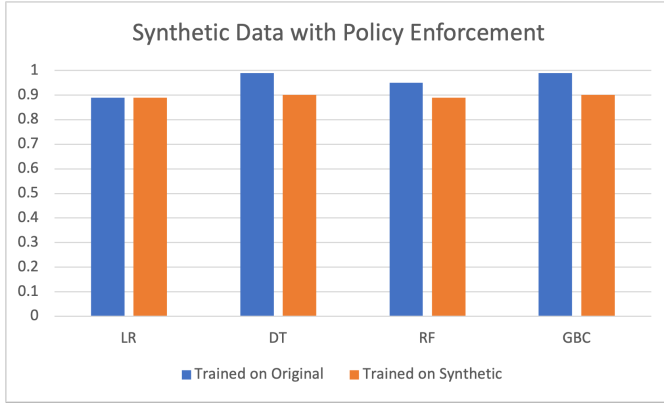
set should be the same to compare the performance of both models. One of the use cases of ML prediction for agricultural data is farmer category prediction. We use 4 ML predictive models here: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Gradient Boosting Classifier (GBC). As observed in Figure 8a and Figure 8b, the ML prediction models have comparable performance on data generated both with and without policy enforcement. The average accuracy of ML models trained on the original data is 0.95. The average accuracy of ML models trained on synthetic data generated without policy enforcement is 0.9. The average accuracy of ML models trained on synthetic data generated without policy enforcement is 0.895. The average loss in accuracy for ML models trained on synthetic data generated without policy enforcement is 0.05. The average loss in accuracy for ML models trained on synthetic data generated without policy enforcement is 0.055. Thus there is an average loss of 0.005 in accuracy of data generated with vs without policy enforcement. Thus the data generated from our framework retains the most relevant information for downstream ML tasks.

D. Privacy Results

1) Attribute Inference Attack: In attribute inference attacks, an attacker tries to reconstruct the original training data or extract sensitive information from the model's parameters. In this setting, an adversary has partial knowledge of some training records and access to a model trained on those records and infers the unknown values of a sensitive feature of those records. Attribute inference attacks typically use ML models to learn about the original dataset from the predictions of the trained model and try to predict the values of unknown sensitive attributes with a confidence score. Efficient Attribute Inference attacks can break privacy-preserving efforts by unmasking the values of sensitive attributes. The lower accuracy of attribute inference attack models proves that privacy-preserving algorithms are efficient in evading these attacks.

(a) Synthetic Data without Policy Enforcement

(b) Synthetic Data with Policy Enforcement

Figure 8: Prediction performance with the model trained on Original vs Privacy-preserving Synthetic Data
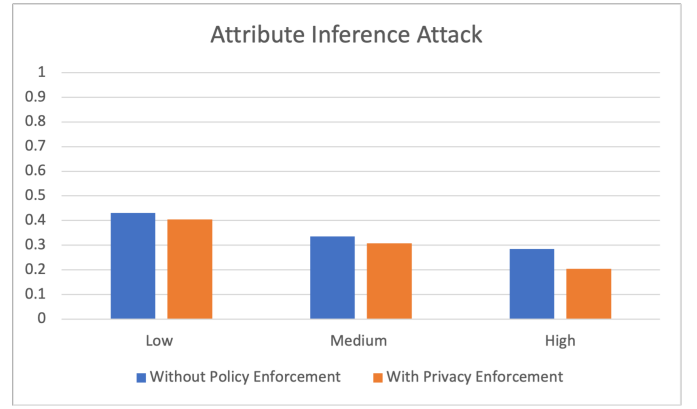

Figure 9: Comparison of Accuracy of Attribute Inference attacks against privacy preserving data generation with vs without policy enforcement
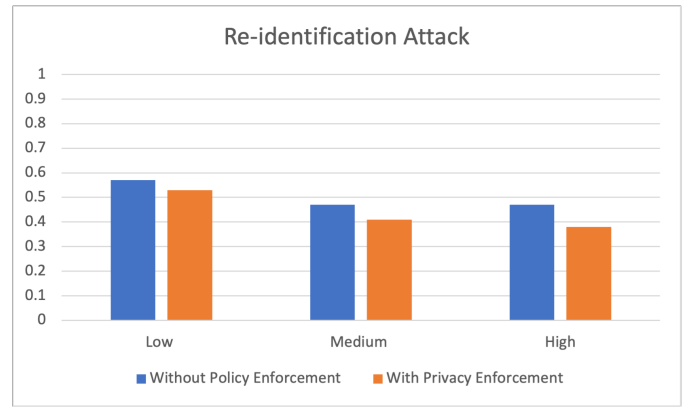

Figure 10: Comparison of Accuracy of Re-identification attacks against privacy preserving data generation with vs without policy enforcement

In Figure 9, we compare the results of a re-identification attack against privacy-preserving datasets generated with and without policy enforcement. For all 3 risk levels, the accuracy of the attribute inference attack model is lower for the dataset generated with our framework of policy enforcement. for privacy-preserving data generated without policy enforcement, the average attribute inference attack accuracy is 0.35 across the 3 privacy risk levels. For privacy-preserving data generated without policy enforcement, the average attribute inference attack accuracy is 0.3 across the 3 privacy risk levels. Thus, there is a 0.05 increase in accuracy loss for data generated with policy enforcement. For attributes with high privacy risk, the increase in accuracy loss is 0.08. This shows that the dataset generated with our approach is more resistant to attribute inference attacks.

2) Re-identification Attack: Re-identification attacks are variations of linkage attacks where even if individual data points are anonymized, attackers could use auxiliary information from other sources to link seemingly anonymous data back to specific individuals, thereby breaching their privacy. Efficient re-identification algorithms use ML to infer information about anonymized datasets with partial knowledge of the original dataset or auxiliary information. It is then used to predict de-anonymized values of sensitive attributes in a dataset. Higher accuracy of re-identification attacks in de-identified datasets shows that the dataset still contains considerable privacy risk. In Figure 10, we compare the results of the re-identification attack against privacy-preserving datasets generated with and without policy enforcement. Data generated without policy enforcement has a consistent accuracy against re-identification attacks for the 3 risk levels of attributes (≈ 0.50). For attributes with low privacy risk, the re-identification attack model has comparable accuracy (0.51) against our framework with policy enforcement. However, for attributes with medium and high privacy risk, our framework with policy enforcement works better with an average accuracy loss of 0.07 in re-identification attacks. Thus the dataset generated with our approach is more resistant against re-identification attacks.

V. RELATED WORK

In the absence of precisely elaborated rules on the appropriate ways to deal with the legal implications of digital

farming technologies on agricultural companies and relationships, farmers and their agribusiness organizations who deliver digital farming technologies started to shape their policies to enhance agricultural data management techniques and provide a foundation for reliable data sharing. Agricultural security and privacy principles and codes of conduct have been developed in different parts of the world. In 2014, in the U.S., the American Farm Bureau's Privacy and Security Principles for Farm Data (2019) [22] was the foremost to draw awareness to some of the problems farmers had about how their information was being collected and shared. After this initiative was 2014 New Zealand's Farm Data Code of Practice [23], and more recently, in 2019, we noticed the E.U. Code of Conduct launch on agricultural data sharing by contractual agreement [3].

With the recent advances in Big Data, there is a gap between the computational capabilities and the data available for research. Data scientists are invested in gathering large volumes of data with secure and privacy-preserving approaches. Privacy-preserving methods and the impact of privacy policies on data sharing have been an ongoing avenue of research [24]–[28]. There has been significant research in designing privacy-preserving data-sharing methods. However, most approaches have a caveat associated with them. Among these various privacy approaches, differential privacy [10], K-anonymity [11], L-diversity [12], and t-closeness [13] are worth noting. It has also been shown that achieving one of the privacy metrics can ensure others [29].

Synthetic data can provide anonymity to original data without loss of accuracy in downstream data analysis tasks. In general, there is a lot of evidence of GANs being used for synthetic data generation and translation in image and text data [14]–[17]. However, the properties of system or device data make it distinct from image and text data. The system data is usually tabular i.e. contains a mix of continuous and discrete variables, and in some cases, the sequence of consecutive rows in the data is important. A conditional generator model can address the issue of mixed attributes in tabular data by seeking to minimize the distance between generated and real data given a fixed value of the discrete variable [8], [30]–[34].

While there are privacy-preserving models for data sharing, these models do not take into account the rules stated in privacy policies such as GDPR [35] [36], EU code of conduct [3] etc. Hence, there is no mechanism to ensure that data shared through these frameworks are compliant with the privacy policies. To bridge this gap, we propose a novel framework that learns rules from the privacy policies and inducts the information in the generation process of a privacy-preserving GAN to ensure that the data generated is privacy-preserving, secure to share, and compliant with the data policies.

Gupta et al. [37] presented a vast exposure to cybersecurity threats and vulnerabilities in smart farming environments. This research [38] proposed PrivySharing, a blockchain-based innovative framework for privacy-preserving and secure IoT data sharing in a smart city environment. West [39] introduced a principles-based framework to assess cyber-attack vulnerabilities and also constructed a precision agriculture system

protected from cyber-attacks. Coble et al. [40] discussed a set of analytical techniques that are increasingly relevant to solving security and privacy issues. Kumar et al. [41] proposed a Secured Privacy-Preserving Framework (SP2F) for smart agricultural Unmanned Aerial Vehicles (UAVs), which handles various cyber attacks.

In addition, several security models for protecting big data in various domains are discussed in [42]–[48].

VI. CONCLUSION AND FUTURE WORK

Technology, particularly the integration of big data analysis, data processing, cloud computing, and IoT devices, plays a pivotal role in enhancing agricultural output both quantitatively and qualitatively. Utilizing these technological advancements enables the agricultural sector to address intricate challenges by effectively analyzing large volumes of crowd-sourced data. However, the sharing of data across various sources raises privacy concerns. While GDPR and similar general privacy regulations provide a foundation for digital data sharing, the specific privacy risks pertinent to agricultural data remain unclear. Defining agricultural big data is still an ongoing debate, making it challenging to delineate privacy concerns. This study delves into the EU Code of Conduct on agricultural data sharing, established through contractual agreements as an extension of GDPR and the E.U. A.I. Act, to comprehensively understand privacy concerns in the realm of agricultural data. Though there have been privacy-preserving efforts in data sharing, there is a lack of policy enforcement in such efforts.

To address the challenge of data sharing while ensuring consistency with privacy regulations, this paper introduces a novel framework for enforcing policy rules in privacy-preserving data sharing. This approach not only mitigates privacy risks like leakage and re-identification but also bridges the gap between privacy-preserving methods and policy-defined privacy requirements. The experimental results show that this framework retains substantial statistical similarity with original data and retains utility in downstream tasks while being resistant to privacy threat models. In conclusion, the proposed framework showcases a promising direction for reconciling agricultural data sharing and privacy concerns, ensuring a sustainable and secure future for agricultural research and advancement. In the future, we would like to extend this framework for policy enforcement in other domains that need privacy-preserving efforts in data sharing, such as health, security, etc.

REFERENCES

[1] H.-L. E. Forum, "How to feed the world in 2050," 2009.
[2] "Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance)."
[3] "Eu code of conduct on agricultural data sharing by contractual agreement," 2020.
[4] EUCommision, "Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts," 2021.

[5] R. Torkzadehmahani, P. Kairouz, and B. Paten, "Dp-cgan: Differentially private synthetic data and label generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[6] J. Xu, X. Ren, J. Lin, and X. Sun, "Dp-gan: diversity-promoting generative adversarial network for generating informative and diversified text," *arXiv preprint arXiv:1802.01345*, 2018.

[7] J. Jordon, J. Yoon, and M. Van Der Schaar, "Pate-gan: Generating synthetic data with differential privacy guarantees," in *International conference on learning representations*, 2018.

[8] A. Kotal, A. Piplai, S. S. L. Chukkapalli, and A. Joshi, "Privetab: Secure and privacy-preserving sharing of tabular data," in *Proceedings of the 2022 ACM on International Workshop on Security and Privacy Analytics*, 2022, pp. 35–45.

[9] "Ceja two new partners sign up to the eu code of conduct on agricultural data sharing, https://www.ceja.eu/two-new-partners-signeu-code-conduct-agricultural-data-sharing/."

[10] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*. Springer, 2006, pp. 265–284.

[11] K. El Emam and F. K. Dankar, "Protecting privacy using k-anonymity," *Journal of the American Medical Informatics Association*, vol. 15, no. 5, pp. 627–637, 2008.

[12] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3–es, 2007.

[13] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 2007, pp. 106–115.

[14] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[16] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5907–5915.

[17] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

[18] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[19] "Modal logic, http://plato.stanford.edu/entries/logic-modal/."

[20] L. Elluri, S. S. L. Chukkapalli, K. P. Joshi, T. Finin, and A. Joshi, "A bert based approach to measure web services policies compliance with gdpr," *IEEE Access*, vol. 9, pp. 148 004–148 016, 2021.

[21] "Itm4impact farmer survey - datasets - ilri datasets portal," https://data.ilri.org/portal/dataset/itm4impact-farmer-survey, (Accessed on 08/30/2023).

[22] A. F. B. Federation, "Privacy and security issues for farm data, centennial," 2019.

[23] "New zealand's farm data code of practice," 2019.

[24] A. Kotal, K. P. Joshi, and A. Joshi, "Vicloud: Measuring vagueness in cloud service privacy policies and terms of services," in *2020 IEEE 13th International Conference on Cloud Computing (CLOUD)*. IEEE, 2020, pp. 71–79.

[25] L. Elluri, K. P. Joshi, and A. Kotal, "Measuring semantic similarity across eu gdpr regulation and cloud privacy policies," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 3963–3978.

[26] A. Kotal, A. Joshi, and K. P. Joshi, "The effect of text ambiguity on creating policy knowledge graphs," in *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*. IEEE, 2021, pp. 1491–1500.

[27] L. Elluri, A. Piplai, A. Kotal, A. Joshi, and K. P. Joshi, "A policy-driven approach to secure extraction of covid-19 data from research papers," *Frontiers in big Data*, vol. 4, p. 701966, 2021.

[28] K. Echenim, L. Elluri, and K. Joshi, "Ensuring privacy policy compliance of wearables with iot regulations," *UMBC Center for Accelerated Real Time Analysis*, 2023.

[29] J. Domingo-Ferrer and J. Soria-Comas, "From t-closeness to differential privacy and vice versa in data anonymization," *Knowledge-Based Systems*, vol. 74, pp. 151–158, 2015.

[30] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 214–223.

[31] L. Xu and K. Veeramachaneni, "Synthesizing tabular data using generative adversarial networks," *arXiv preprint arXiv:1811.11264*, 2018.

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

[33] N. Das, A. Kotal, D. Roseberry, and A. Joshi, "Change management using generative modeling on digital twins," in *2023 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 2023, pp. 1–6.

[34] A. Piplai, A. Kotal, S. Mohseni, M. Gaur, S. Mittal, and A. Joshi, "Knowledge-enhanced neurosymbolic artificial intelligence for cybersecurity and privacy," *IEEE Internet Computing*, vol. 27, no. 5, pp. 43–48, 2023.

[35] L. Elluri, A. Nagar, and K. P. Joshi, "An integrated knowledge graph to automate gdpr and pci dss compliance," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 1266–1271.

[36] L. Elluri and K. P. Joshi, "A knowledge representation of cloud data controls for eu gdpr compliance," in *2018 IEEE World Congress on Services (SERVICES)*. IEEE, 2018, pp. 45–46.

[37] M. Gupta, M. Abdelsalam, S. Khorsandroo, and S. Mittal, "Security and privacy in smart farming: Challenges and opportunities," *IEEE access*, vol. 8, pp. 34 564–34 584, 2020.

[38] I. Makhdoom, I. Zhou, M. Abolhasan, J. Lipman, and W. Ni, "Privysharing: A blockchain-based framework for privacy-preserving and secure data sharing in smart cities," *Computers & Security*, vol. 88, p. 101653, 2020.

[39] J. West, "A prediction model framework for cyber-attacks to precision agriculture technologies," *Journal of Agricultural & Food Information*, vol. 19, no. 4, pp. 307–330, 2018.

[40] K. H. Coble, A. K. Mishra, S. Ferrell, and T. Griffin, "Big data in agriculture: A challenge for the future," *Applied Economic Perspectives and Policy*, vol. 40, no. 1, pp. 79–96, 2018.

[41] R. Kumar, P. Kumar, R. Tripathi, G. P. Gupta, T. R. Gadekallu, and G. Srivastava, "Sp2f: A secured privacy-preserving framework for smart agricultural unmanned aerial vehicles," *Computer Networks*, vol. 187, p. 107819, 2021.

[42] D. Gupta, S. Bhatt, M. Gupta, O. Kayode, and A. S. Tosun, "Access control model for google cloud iot," in *2020 IEEE 6th Intl conference on big data security on cloud (BigDataSecurity), IEEE Intl conference on high performance and smart computing,(HPSC) and IEEE Intl conference on intelligent data and security (IDS)*. IEEE, 2020, pp. 198–208.

[43] D. Gupta, S. Bhatt, M. Gupta, and A. S. Tosun, "Future smart connected communities to fight covid-19 outbreak," *Internet of Things*, vol. 13, p. 100342, 2021.

[44] Ö. Aslan, M. Ozkan-Okay, and D. Gupta, "Intelligent behavior-based malware detection system on cloud computing environment," *IEEE Access*, vol. 9, pp. 83 252–83 271, 2021.

[45] M. Ozkan-Okay, R. Samet, Ö. Aslan, and D. Gupta, "A comprehensive systematic literature review on intrusion detection systems," *IEEE Access*, vol. 9, pp. 157 727–157 760, 2021.

[46] D. Gupta, M. Gupta, S. Bhatt, and A. S. Tosun, "Detecting anomalous user behavior in remote patient monitoring," in *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 2021, pp. 33–40.

[47] D. Gupta, O. Kayode, S. Bhatt, M. Gupta, and A. S. Tosun, "Hierarchical federated learning based anomaly detection using digital twins for smart healthcare," in *2021 IEEE 7th International Conference on Collaboration and Internet Computing (CIC)*. IEEE, 2021, pp. 16–25.

[48] D. Gupta, S. Bhatt, P. Bhatt, M. Gupta, and A. S. Tosun, "Game theory based privacy preserving approach for collaborative deep learning in iot," in *Deep Learning for Security and Privacy Preservation in IoT*. Springer, 2022, pp. 127–149.