IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 January 2025, accepted 21 February 2025, date of publication 28 February 2025, date of current version 10 March 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3546618

SURVEY

Enhancing Data Privacy: A Comprehensive Survey of Privacy-Enabling Technologies

QAISER RAZI[1], RAJA PIYUSH[2], ARJAB CHAKRABARTI[2], ANUSHKA SINGH[2], VIKAS HASSIJA[2], AND G. S. S. CHALAPATHI[1], (Senior Member, IEEE)

[1]Department of Electrical and Electronics Engineering, Birla Institute of Technology and Science (BITS) Pilani, Pilani Campus, Rajasthan 333031, India
[2]School of Computer Engineering, Kalinga Institute of Industrial Technology (KIIT), Bhubaneswar, Odisha 75102, India

Corresponding author: G. S. S. Chalapathi (gssc@pilani.bits-pilani.ac.in)

This work was supported by the Birla Institute of Technology and Science, Pilani (BITS Pilani).

ABSTRACT Privacy is a fundamental human right, especially crucial in our modern digital age. With the rapid advancement of technology, ensuring individuals' privacy has become increasingly complex. Our survey paper aims to shed light on various privacy engineering technologies that play a crucial role in protecting personal data. We delve into four key areas: data anonymization, data encryption, synthetic data generation, and differential privacy. These technologies serve as essential tools in safeguarding online privacy. Data anonymization, for instance, includes removing or modifying identifiable information from datasets to protect individuals' identities. Encryption secures data by converting it into a code that can only be decoded by authorized parties. Synthetic data generation creates artificial data that closely resembles real data but doesn't contain any identifiable information. Differential privacy adds a small amount of controlled noise to protect sensitive information. Throughout our exploration, we not only explain the principles and techniques behind these technologies but also the tools used for each of these techniques and evaluation criteria and also examine their practical applications. By understanding their strengths, limitations, and real-world implementations, we gain valuable insights into how they contribute to the broader goal of ensuring privacy in our digital world.

INDEX TERMS Privacy engineering, data anonymization, data encryption, synthetic data, differential privacy, privacy preservation, privacy technologies.

I. INTRODUCTION

In the digital age, where data is ubiquitously generated, collected, and analyzed, the need for robust data privacy cannot be overstated [1]. The advent of sophisticated technologies has revolutionized the way we interact with data, rendering traditional privacy methods inadequate [2]. This paper delves into the evolving landscape of privacy engineering, a discipline that integrates privacy into the fabric of technology design and deployment [3]. The unprecedented growth of the Internet, along with the rise of big data analytics and the Internet of Things (IoT), has led to the generation of vast amounts of personal data [4]. This surge in data production has dramatically increased the potential for privacy breaches, making data privacy not just a luxury but a necessity. In today's interconnected world, protecting individual privacy is of utmost importance, as personal data can be highly sensitive, and its misuse can have far-reaching consequences [5]. Contemporary data ecosystems are fraught with numerous threats that challenge the privacy of individuals. These threats range from sophisticated cyber-attacks targeting personal data stored in cloud services to pervasive surveillance practices that monitor online activities [6]. Additionally, the advent of technologies such as Artificial Intelligence (AI) and Machine Learning (ML) has intensified the risk of data breaches, as these technologies can analyze data at an unprecedented scale, often without the knowledge or consent of authors [7]. Historically, privacy protection in the digital realm has relied heavily on encryption, access control mechanisms, and data anonymization techniques [8]. While these methods provided a foundational layer of security, they often fell short

The associate editor coordinating the review of this manuscript and approving it for publication was Zijian Zhang.

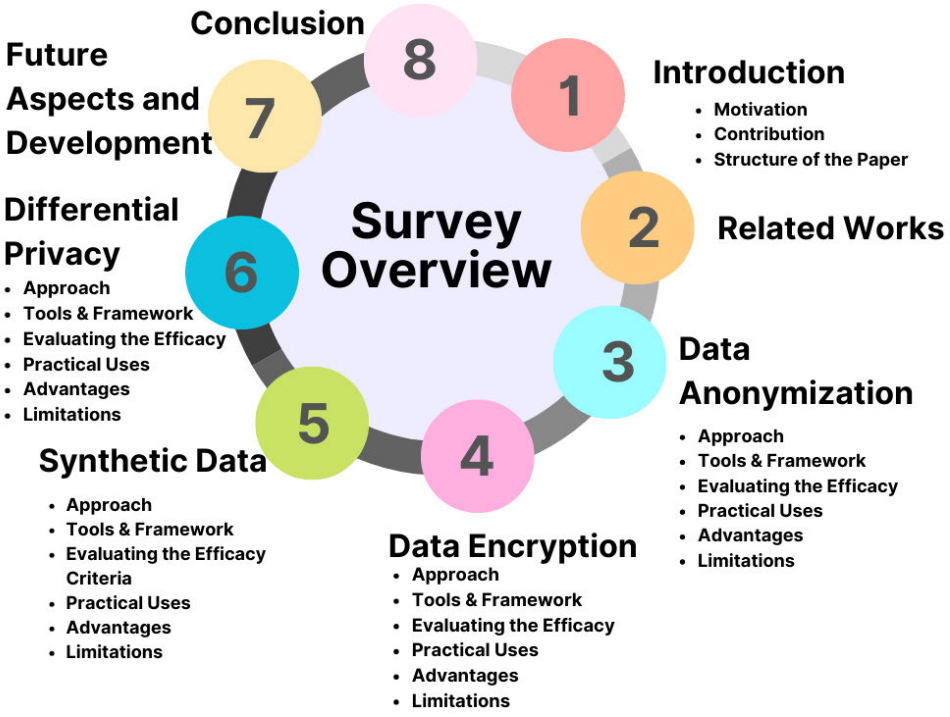

FIGURE 1. Survey overview.

in the face of rapidly advancing technological capabilities and increasingly sophisticated cyber threats [9]. The limitations of these traditional approaches have paved the way for the emergence of privacy engineering as a distinct and critical field in the tech industry [10]. This paper will explore the evolution of privacy engineering technologies, highlighting how they are reshaping the approach to data privacy and are offering a holistic solution to the ''privacy for one as privacy for all'' paradigm.

A. MOTIVATION

The motivation for exploring privacy engineering arises from the critical need to address privacy challenges posed by the exponential growth of data generation and the accompanying risks. Traditional methods like encryption and anonymization are increasingly insufficient against modern threats, particularly with the rise of big data, IoT, AI, and ML. Recent advances in IoT, AI, and ML have allowed the scientific community to efficiently mine vast amounts of information worldwide and to extract new knowledge by discovering hidden patterns and correlations. Nevertheless, all this shared information can be used to invade the privacy of individuals through the use of fusion and mining techniques. Simply removing direct identifiers such as name, or phone number is not anymore sufficient to safeguard privacy. In numerous cases, other fields, like gender, date of birth, zip code, and even usage patterns or choices of services, can be used to re-identify individuals and to expose their sensitive details, e.g., their medical conditions, financial statuses, and transactions, or even their private connections [11]. Some adversarial knowledge is aggregated with the help of AI techniques and subsequently used to compromise the privacy of individuals [12]. In 2015, Google DeepMind partnered with the UK's National Health Service (NHS) to develop an app called Streams. It was designed to detect acute kidney injury. However, it was later found that 1.6 million patient records were shared with DeepMind without patient consent [13].

These technologies amplify the potential for privacy breaches, making robust privacy integration essential to protect sensitive information and maintain trust in digital interactions. The scope of this work is to provide an in-depth overview of the current state of the art in data privacy. This paper aims to highlight innovative privacy engineering approaches that provide comprehensive protection and adapt to evolving technological landscapes.

B. OUR CONTRIBUTION

The contributions of this paper are multifaceted, providing an in-depth analysis of key privacy engineering technologies and their benefits. Our work delves deeply into data anonymization, data encryption, and synthetic data generation, offering comprehensive insights into each method's technical aspects and practical applications. This detailed exploration is a useful resource for understanding the nuances and advantages of integrating privacy measures into technology design. The main contribution of our paper is listed below:

- Data Anonymization:
  1) Examined various anonymization techniques such as k-anonymity, l-diversity, and t-closeness.

2) Presented the effectiveness of these techniques in protecting against re-identification attacks.
3) Discussed the trade-offs between data utility and privacy.

- Data Encryption:
  1) Explored symmetric and asymmetric encryption methods.
  2) Assessed the security and performance implications of different encryption algorithms.
  3) Provided guidelines for selecting appropriate encryption techniques based on data sensitivity and application requirements.
- Synthetic Data Generation:
  1) Investigated methods for generating synthetic data that mimics real datasets.
  2) Discussed the use of synthetic data in preserving privacy while maintaining data utility for analysis.
  3) Highlighted the challenges and limitations of synthetic data generation, including the risk of model inversion attacks.
- Differential Privacy:
  1) Discussed various differential privacy methods.
  2) Provided various tools and frameworks to achieve differential privacy.
  3) Discussed various applications of differential privacy, including its advantages and limitations.

C. METHODOLOGY FOR SELECTING PAPERS SURVEYED IN THIS WORK

We have included seminal works in each domain discussed in this survey, i.e., data anonymization, encryption, synthetic data, and differential privacy. Apart from these works, we have given more focus to discussing the techniques adopted in recent years [2017-2024] because these domains have changed rapidly with the proliferation of machine learning and artificial intelligence technologies. After gathering the papers published in the aforementioned years, we have filtered them based on the venues in which they are published (high-quality peer-reviewed conferences, high-impact factors journals, leading technical books, and online articles), their citation count, etc.

D. STRUCTURE OF THE PAPER

The paper is structured as follows: Section II introduces the background work done in this field. Section III discusses data anonymization, including its approach, relevant tools, implementation considerations, evaluation metrics, and real-world applications. Section IV covers data encryption techniques, their implementation methodologies, evaluation criteria, practical applications, advantages, and limitations. Synthetic data generation techniques are investigated in Section V, offering insights into the approach, tools and frameworks, implementation strategies, evaluation criteria, practical uses, advantages, and limitations. Section VI discusses differential privacy, its technical background, various tools and frameworks, its evaluation criteria, practical use cases, applications, and limitations. Section VII discusses the future aspects and emerging trends in privacy engineering technologies. Finally, Section VIII summarizes the lessons learned and concludes the paper. The organizational overview of this survey is also shown in Fig. 1.

II. RELATED WORKS

The field of privacy engineering has witnessed significant advancements in recent years, driven by the escalating concerns over data privacy and the proliferation of sensitive information across large domains. Fig. 2 illustrates the interplay between evolving technologies, artificial intelligence/machine learning (AI/ML), and privacy engineering. It underscores how the Internet of Things (IoT), Data Analytics, and AI/ML contribute to privacy engineering, ensuring encryption, data fortification, access controls, managing data entry points, and personal privacy protection. This section presents a comprehensive review of related works in privacy engineering, focusing on techniques and technologies to safeguard individuals' privacy while enabling effective data utilization as summarized in Table 1.

Spiekermann and Cranor [14] presented a comprehensive overview of privacy engineering by integrating insights from various research areas. It begins by discussing privacy requirements from historical and contemporary perspectives, using a three-layer model to link user privacy concerns to system operations. In the second part, guidelines for building privacy-friendly systems are developed, distinguishing between "privacy-by-policy" and "privacy-by-architecture" approaches. The former emphasizes implementing fair information practices, while the latter focuses on minimizing identifiable personal data collection and prioritizing anonymization and client-side data processing. The privacy-by-policy approach involves setting and following rules to manage personal data responsibly, ensuring compliance with laws like the General Data Protection Regulation (GDPR) or the California Consumer Privacy Act (CCPA). It emphasizes transparency by informing users about how their data is collected and used, typically through consent forms and privacy policies. Users are given control over their data, such as the ability to view, update, or delete it. However, this method can be difficult for users to navigate due to complex legal language, which may lead to confusion or consent fatigue [19], [20]. On the other hand, the privacy-by-architecture approach incorporates privacy protections directly into the design of systems. It focuses on minimizing the collection of personal data by using techniques like anonymization and processing data locally on user devices. While this approach provides stronger and more reliable privacy protections without depending on users understanding legal policies, it can be more complex and expensive to implement [21], [22].

Boreale et al. [15] examines group-based anonymization schemes, a prevalent method in data publishing for safeguarding individuals' privacy. These schemes release

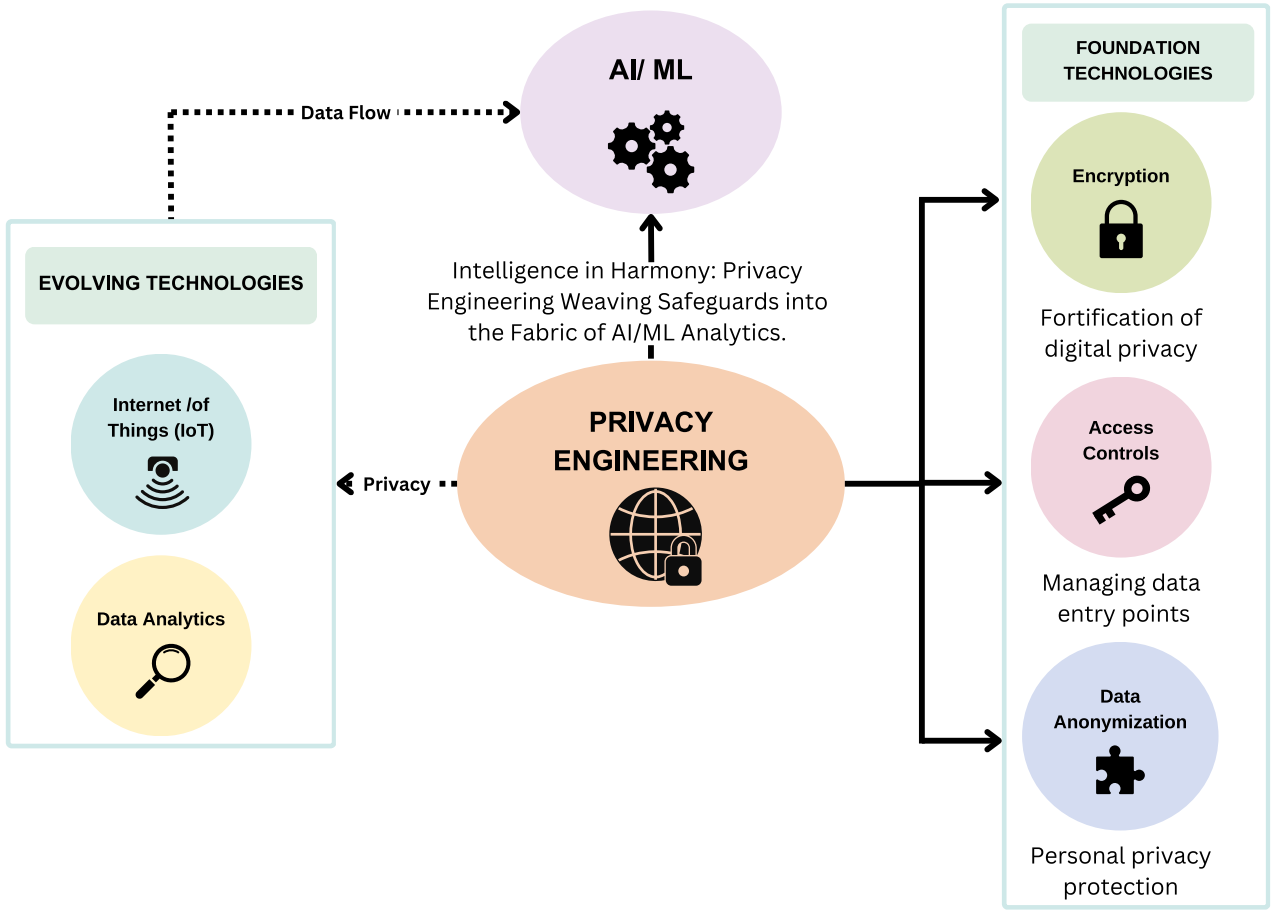

FIGURE 2. Privacy engineering revolution.

TABLE 1. Related surveys in privacy engineering.

Reference	Description	Contributions	Limitations
Spiekermann and Cranor [14]	Comprehensive overview of privacy engineering, integrating insights from various areas. Discusses privacy requirements, system operations, and guidelines for building privacy-friendly systems.	Holistic understanding of privacy engineering. Practical guidelines for implementation.	Further validation and adaptation are needed in real-world contexts.
Boreale *et al.* [15]	Examines group-based anonymization schemes for safeguarding privacy. Proposes Bayesian model and Monte Carlo methodology for risk analysis.	Systematic approach for assessing privacy risks. Demonstrates effectiveness through real-world analysis.	Requires expertise in Bayesian methods for implementation.
Dos Santos *et al.* [16]	Presents a synthetic data generation system using variational autoencoder (VAE) and linked data paradigm. Demonstrates applicability in energy management.	Facilitates expansion of datasets. Demonstrates potential for real-world implementation.	Requires further validation across diverse domains.
Gai *et al.* [17]	Addresses privacy concerns in cloud computing. Proposes Dynamic Data Encryption Strategy (D2ES) for selective encryption.	Novel approach to enhancing privacy. Empirical validation demonstrates efficacy.	Implementation complexity may hinder adoption.
Yadav and Bondre [18]	Explores secure data transmission in computer networks. Proposes encryption algorithm for online transactions.	Enhances security, integrity, and authentication. Integrates multiple cryptographic techniques.	Practical deployment may require compatibility considerations.
Our Work	**Provides an in-depth overview of data anonymization, data encryption, synthetic data generation, and differential privacy in detail**.	**Provides comprehensive insights into privacy engineering technologies and how they are beneficial.**	-

an obscured version of the original data, concealing the precise associations between individuals and attribute values. Balancing data utility with the risk posed by potential attackers targeting individuals in the dataset is crucial. They

propose a unified Bayesian model of group-based schemes and an associated Markov Chain Monte Carlo (MCMC) methodology for learning population parameters from an anonymized table to address this. This technique facilitates the analysis of the risk of individuals in the dataset being linked to specific sensitive values, given knowledge of their nonsensitive attributes. This analysis, termed relative threat analysis, is demonstrated using a real-world dataset, showcasing the efficacy of the proposed methodology. Regarding methodology, Bayesian modeling, and MCMC sampling are complementary tools in privacy-preserving data analysis. Bayesian models are useful for incorporating prior beliefs or knowledge into the analysis and adjusting those beliefs based on the observed data. This makes the Bayesian approach highly adaptable in real-world settings where prior information may be available. On the other hand, MCMC is essential for making Bayesian inference feasible in complex models, particularly when the posterior distribution is analytically intractable [23]. Bayesian inference, by itself, offers a structured way to estimate probabilities and can be particularly helpful for evaluating uncertainty in anonymized data. MCMC, however, is critical when the dataset is too complex for simple analytical solutions, enabling the model to generate approximate solutions through sampling.

Together, these two approaches create a robust framework for analyzing privacy risks in anonymized data, particularly when attempting to balance privacy with the usefulness of the data. By leveraging Bayesian models with MCMC, researchers can estimate the risk of re-identification and optimize anonymization techniques to better protect individuals' privacy.

Dos Santos et al. [16] presents a synthetic data generation system utilizing the variational autoencoder (VAE) technique and the linked data paradigm to expand small datasets into larger ones, acting as representative samples. The proposed approach involves a Linked Data-based dataset extractor to obtain samples in a given context, followed by training a VAE on these samples to learn the latent distribution characterizing the dataset. This enables the generation of new records resembling those in the original dataset. The paper includes several case studies in energy management, demonstrating the process's applicability and efficiency in automating data generation for smart energy management. Results indicate the system's capability to produce synthetic datasets across diverse energy contexts, showcasing its potential for real-world implementation. The VAE technique stands out when compared to other traditional synthetic data generation methods because it has the ability to model the latent structure of the data. Compared to other generative models like Generative Adversarial Networks (GANs), VAE has the advantage of being more stable and easier to train, as it doesn't require the adversarial training process found in GANs. While GANs are known for the production of highly realistic synthetic data, they can still suffer from training instability and mode collapse, whereas in the case of VAE, the model generates only a limited variety of outputs, ensuring a more comprehensive exploration of the latent space, allowing for a broader range of synthetic data generation. Gai et al. [17] addresses the escalating privacy concerns accompanying the proliferation of big data applications in cloud computing. While acknowledging the transformative impact of these technologies on service models and application performance, they highlight the challenges posed by the exponential increase in data volume. Specifically, they identify the prolonged execution time of data encryption as a critical issue during processing and transmission, often leading to the compromise of privacy for the sake of performance optimization. The authors propose a novel Dynamic Data Encryption Strategy (D2ES) approach to tackle this issue. Their method focuses on selectively encrypting data and employing privacy classification methods within predefined timing constraints. By prioritizing data privacy while adhering to execution time requirements, D2ES maximizes the scope of privacy protection. The authors validate the efficacy of D2ES through comprehensive experiments, providing empirical evidence of its effectiveness in enhancing privacy in practical settings. Unlike traditional encryption techniques that uniformly apply the same level of encryption to all data, regardless of its sensitivity, D2ES adopts a more refined and adaptive approach. Conventional methods often consume significant resources, leading to slower processing speeds and reduced overall system efficiency as data volumes increase. In contrast, D2ES streamlines the encryption process by prioritizing the protection of sensitive information through robust encryption measures, while applying lighter security or bypassing encryption for less critical data [20], [24]. This approach achieves a balance between maintaining privacy and optimizing performance. In traditional systems, encrypting large amounts of data can extend processing times, delaying service delivery. D2ES addresses this by dynamically adjusting encryption levels based on the data's sensitivity and system timing requirements. This ensures that performance benchmarks are met while safeguarding privacy, even as data volumes grow [25], [26]. D2ES provides a strategic solution to the growing demand for privacy in big data ecosystems, where conventional encryption strategies struggle to keep up with exponentially increasing data. By incorporating adaptability and timing awareness, D2ES ensures an optimal balance between privacy and system performance, making it particularly effective for cloud-based applications.

Yadav and Bondre [18] explored computer networks, emphasizing secure data transmission's vital role. They propose a novel encryption algorithm for online transactions, combining private and public key cryptography techniques. This algorithm integrates AES for encryption, Dual-RSA for authentication, Elliptic Curve Digital Signature Algorithm (ECDSA) for digital signatures, SHA-1 for integrity, and Elliptic-curve Diffie-Hellman (ECDH) for public key generation. The Advanced Encryption Standard (AES) is a

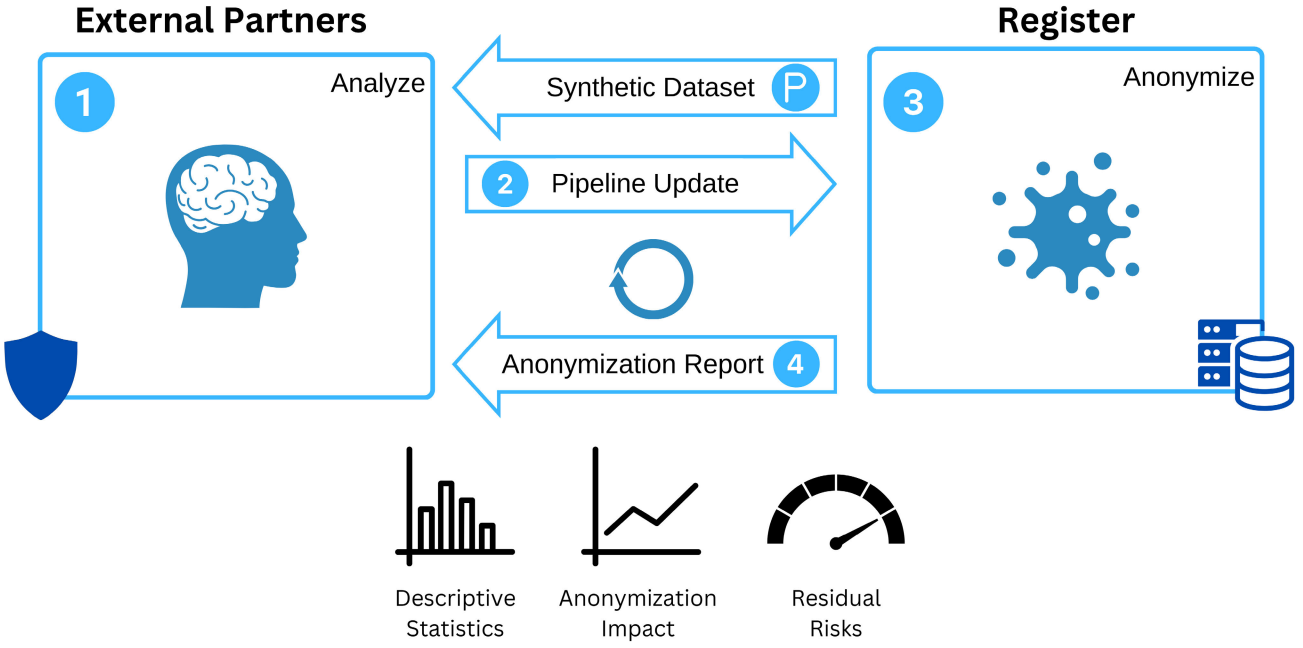

FIGURE 3. Data Anonimyzation Overview.

highly popular symmetric encryption algorithm used for safeguarding data in various applications, including personal data exchange and image encryption. It processes data in 128-bit blocks and enhances security through substitution-permutation operations, key expansion, and S-box mechanisms, making it resistant to cryptographic attacks. AES is versatile and efficiently implemented in hardware, such as Field Programmable Gate Array (FPGA) designs, and has been combined with chaotic maps to strengthen image encryption techniques. It provides a reliable solution for ensuring data confidentiality and integrity in modern digital systems. Combining Dual-RSA for authentication, ECDSA for digital signatures, SHA-1 for data integrity, and ECDH for public key exchange creates a secure framework for digital communication. Dual-RSA uses separate Rivest-Shamir-Adleman (RSA) keys for encryption and authentication, increasing security levels. ECDSA enables efficient digital signatures with robust protection, particularly useful in resource-limited environments. Although SHA-1 is now considered less secure, it has historically been used to verify data integrity. ECDH facilitates secure key exchange, allowing parties to establish a shared secret over unsecured networks. Together, these technologies form a strong and comprehensive approach to securing digital communication systems. Their approach aims to enhance data transmission security, integrity, and authentication.

III. DATA ANONYMIZATION

A. OVERVIEW

In an era where personal data is both a valuable resource and a vulnerability, data anonymization emerges as a pivotal strategy in safeguarding individual privacy [27]. This technique revolves around concealing personal identifiers in a dataset. It is increasingly relevant in a world brimming with data-driven decisions. As organizations and governments collect vast amounts of personal information, the ethical and legal imperatives to protect this data are more pronounced than ever [28]. Data anonymization does not just serve compliance with privacy laws like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA); it also embodies a commitment to ethical data practices. The core principle of data anonymization is the disassociation of data from identifiable individuals [29]. This involves transforming personal data so that individuals cannot be directly or indirectly identified by the data, ensuring privacy while retaining the data's utility. Fig. 3 represents the steps of the data management process involving three steps: analysis by external partners, pipeline updates with synthetic datasets and anonymization reports, and registration. It also highlights the outcomes of this process: descriptive statistics, anonymization impact, and residual risks. The diagram effectively visualizes complex data handling processes and their results. The significance of this balance cannot be overstated, as the value of data often lies in its ability to inform and guide decisions. However, with the rising sophistication of data analysis techniques and the burgeoning volume of data collected, the challenges of effective anonymization are increasingly complex.

Anonymization techniques have evolved from basic methods like removing direct identifiers to more advanced statistical and algorithmic approaches considering the broader context of data usage [30]. These advancements are crucial in an environment where re-identification risks are ever-present and the consequences of privacy breaches can be severe.

TABLE 2. Comparison of data anonymization techniques.

Technique	Description	Advantages	Disadvantages
Generalization [39]	Replacing specific values with broader categories.	Preserves privacy, retains some utility.	Potential loss of data granularity, risk of identification.
Data Masking [40]	Substituting sensitive data with fictitious or scrambled values.	Simple implementation preserves data format.	Limited effectiveness against sophisticated attacks, potential for reverse engineering.
Perturbation [32]	Introducing slight changes to data values while preserving statistical properties.	Preserves statistical properties and retains utility.	Risk of data distortion, the potential for inference attacks.
Data Swapping [41]	Data anonymization technique where values of specific attributes are swapped between records in a dataset, ensuring that the relationships between data points are maintained, but individual records are anonymized.	Maintains statistical properties of the dataset while anonymizing individuals.	Can become complex if too many attributes are swapped.
Pseudonymization [42]	Replaces identifiable data with pseudonyms or tokens.	Balances privacy and data utility, allowing re-identification when necessary.	Poses a risk of re-identification if the mapping key is compromised, offering less security than full anonymization.

As such, understanding and implementing effective data anonymization practices have become indispensable skills for privacy engineers and data scientists alike [31]. The technical underpinnings of data anonymization are deeply rooted in the principles of statistical science and computer security. The primary objective of these techniques is to modify personal data in such a way that the identity of individuals cannot be ascertained, directly or indirectly, while maintaining the data's usefulness for analysis and decision-making [32].

B. TECHNICAL BACKGROUND

The primary objective of these techniques is to modify personal data in such a way that the identity of individuals cannot be ascertained, directly or indirectly, while maintaining the data's usefulness for analysis and decision-making. Table 2 compares various data anonymization techniques. It includes Generalization, which replaces specific values with broader categories; Data Masking, which substitutes sensitive data with fictitious or scrambled values; Perturbation, which introduces slight changes to data values; Data swapping in which attribute values in a dataset are rearranged; Differential Privacy, which adds controlled noise to data; and Pseudonymization, where personal identifiers are replaced with pseudonyms. The working of anonymization is explained below.

1) Generalization: This involves abstracting personal data by reducing its precision [33]. For example, they can replace exact ages with ages grouped in ranges or precise locations with broader geographical areas. Generalization helps protect individual identity but can reduce the granularity and, hence, the utility of the data. This method is particularly effective in ensuring compliance with privacy standards like HIPAA in healthcare or GDPR in general data protection. However, overly broad generalization can severely impact the utility of datasets for machine learning or data analytics tasks.
2) Perturbation: This technique involves adding 'noise' to the data – slight alterations that prevent exact values from being known yet preserve statistical properties [32]. The challenge here lies in determining the right amount of noise to ensure privacy without significantly compromising data accuracy. Perturbation is often used in differential privacy frameworks to introduce randomness, enabling privacy-preserving statistical analysis and machine learning on sensitive datasets.
3) K-Anonymity: A foundational model in data anonymization, Each record is guaranteed to be indistinguishable from a minimum of $k - 1$ other records using k-anonymity with respect to certain identifying attributes [34]. This model's strength is in creating ambiguity in identification, though it has limitations in dealing with high-dimensional data and does not account for sensitive attribute disclosure [35]. Despite its wide adoption, implementing k-anonymity often requires careful tuning to ensure both sufficient privacy and minimal loss of data utility.
4) L-Diversity: An extension of k-anonymity, l-diversity requires that each equivalence class has at least 'l' well-represented values for sensitive attributes [36]. This method addresses some limitations of k-anonymity by protecting against attribute disclosure attacks but can still be susceptible to skewness and similarity attacks [37]. L-diversity is particularly effective in mitigating risks in datasets with categorical sensitive attributes, such as patient medical records or demographic surveys.
5) T-Closeness: Further advancing the concept, t-closeness demands that the distribution of a sensitive attribute in any equivalence class is not more than a threshold 't' different from the distribution of the attribute in the overall dataset [21]. This approach seeks to maintain the data's statistical properties more effectively. However, achieving t-closeness often

comes at the expense of significant data suppression or generalization, which may impact analysis accuracy in practice.

1) APPROACH TO DATA ANONYMIZATION

The approach to data anonymization is a meticulous process that necessitates a deep understanding of the data at hand and the overarching privacy objectives. This process involves several key steps:

1) **Identifying Sensitive Data:**
   a) Data Classification: The first step is to classify the data, distinguishing between sensitive and non-sensitive information [38]. This classification is crucial as it dictates which data elements require anonymization.
   b) Contextual Analysis: Understanding the context in which the data will be used is essential. Different contexts may require different levels of privacy protection, influencing the anonymization approach.
2) **Risk Assessment:**
   a) Threat Modeling: Identifying potential threats to privacy, such as the likelihood of data re-identification, is a critical component [43]. This involves understanding the capabilities of potential adversaries and the methods they might use to breach privacy.
   b) Privacy Impact Assessment: Conducting a comprehensive assessment to understand the impact on privacy if the data were to be compromised. This helps determine the level of anonymization needed.
3) **Balancing Privacy and Utility:**
   a) Determining Anonymization Level: Deciding how much to anonymize the data is a delicate balance [44]. Over-anonymization can render the data useless, while under-anonymization can leave privacy vulnerabilities.
   b) Iterative Process: Often, achieving the right balance requires an iterative process, where anonymization levels are adjusted based on feedback and testing.

2) TOOLS & FRAMEWORKS

The implementation of data anonymization strategies heavily relies on specialized tools and frameworks. These are designed to offer robust, efficient, and user-friendly methods for transforming sensitive data into anonymized formats. Here, we explore some key tools and frameworks, their features, and how they cater to different anonymization needs.

1) **ARX Data Anonymization Tool [19]:**
   a) Principle of Operation: ARX functions by utilizing privacy models such as k-anonymity, l-diversity, and t-closeness for the transformation of datasets. It applies methods like generalization, suppression, and data perturbation to modify sensitive information, reducing the risk of re-identification. The tool continuously evaluates privacy risks and allows users to adjust anonymization parameters interactively. ARX also provides real-time feedback, helping users visualize the trade-offs between maintaining data utility and enhancing privacy, ensuring a balanced approach to data protection [45].
   b) Features: ARX is a comprehensive tool that supports a wide range of anonymization techniques, including k-anonymity, l-diversity, and t-closeness. It's equipped with a user-friendly interface and provides visual aids to help users understand the impact of different anonymization settings.
   c) Use Cases: Ideal for researchers and organizations that need a versatile tool for exploring various anonymization strategies. It is particularly useful in scenarios where different levels of data sensitivity require a flexible approach.
2) **sdcMicro [46]:**
   a) Principle of Operation: sdcMicro works by employing techniques like data perturbation, which slightly modifies individual data points to obscure personal identifiers while retaining statistical accuracy. It also utilizes micro-aggregation, which groups similar records and replaces individual values with group averages, masking individual information. Variable reduction is another method applied, where highly identifying variables are either generalized or suppressed. The tool offers risk assessment modules to calculate the risk of re-identification, enabling users to fine-tune anonymization settings according to privacy requirements.
   b) Features: This R package specializes in microdata anonymization, offering methods like data perturbation, variable reduction, and micro-aggregation. It is particularly known for its robustness in statistical data anonymization.
   c) Use Cases: sdcMicro is well-suited for statisticians and data analysts working with survey data or any form of microdata where maintaining statistical properties post-anonymization is crucial.
3) **Microsoft's Presidio [47]:**
   a) Principle of Operation: Presido operation involves scanning and identifying personal data within the text by using machine learning algorithms and predefined regular expressions. Once detected, sensitive information can be either replaced with generic terms or generalized to a less

specific form. This automated process allows for efficient anonymization of unstructured text while maintaining data utility for further analysis.
b) Features: Presidio focuses on identifying and anonymizing sensitive text data. It uses machine learning models to detect various types of personal information and offers options for anonymization, including replacement and generalization.
c) Use Cases: It is useful for businesses and developers handling large volumes of unstructured text data, such as customer feedback, where personal information needs to be identified and anonymized efficiently.

4) **IBM Guardium** [48]:
a) Principle of Operation: IBM Guardium is an advanced data security solution to detect and prevent unauthorized database activities. It uses anomaly detection techniques to identify deviations in user behavior, relying on both self-consistency and global consistency checks to assess whether actions align with normal behavior patterns. Guardium also incorporates machine learning models trained on historical data to monitor database access, providing real-time alerts for suspicious activities. Its modular design allows seamless integration into existing database architectures, ensuring continuous monitoring and protection.
b) Features: IBM Guardium offers comprehensive features, including real-time monitoring, advanced anomaly detection, data classification, and compliance reporting. It provides visualization tools to investigate alerts, user activity profiling, and feedback loops to refine its detection algorithms.
c) Use Cases: This tool is ideal for industries such as finance, healthcare, and telecommunications that require robust database security. It effectively addresses insider threats, stolen credentials, and policy violations, making it a critical asset for organizations managing sensitive data.

3) EVALUATING ANONYMIZATION EFFICACY

1) **Privacy Protection [9]:**
a) Re-identification Risk Assessment: Evaluating the risk of re-identifying individuals in the anonymized dataset. This involves statistical tests to estimate the likelihood that anonymized data can be linked back to individuals.
b) Adversarial Testing: Simulating potential attacks to test the robustness of the anonymization. This can involve attempts to de-anonymize data using available external information.

2) **Data Integrity [49]:**
a) Statistical Analysis: Ensuring the anonymized data retains essential statistical properties. This includes comparing the original and anonymized datasets distributions, means, and variances.
b) Use Case Validation: This involves testing the anonymized data against specific use cases to ensure it provides meaningful insights and supports decision-making processes.

3) **Compliance with Regulations [50]:**
a) Legal and Regulatory Adherence: Verifying that the anonymization process fully complies with relevant data protection laws, such as GDPR (General Data Protection Regulation), HIPAA (Health Insurance Portability and Accountability Act), or CCPA (Central Consumer Protection Authority). This often involves a legal review and documentation process.
b) Audit Trails and Documentation: This is achieved by maintaining comprehensive records of the anonymization process, including decisions made and methodologies applied, to demonstrate compliance and facilitate audits.

4) **Performance Metrics [28]:**
a) Efficiency Analysis: Evaluating the computational efficiency of the anonymization process is particularly important in large-scale data operations. This includes measuring the processing time and resource utilization.
b) Scalability Assessment: Testing the scalability of anonymization methods, ensuring they remain effective as data volumes increase.

5) **User Feedback and Stakeholder Satisfaction [51]:**
a) Feedback from End-Users: Collect feedback from those who utilize the anonymized data, such as data scientists, analysts, and decision-makers, to assess the utility and applicability of the data.
b) Stakeholder Review: Engaging with various stakeholders, including data privacy officers and legal teams, to ensure that the anonymization meets organizational and regulatory expectations.

C. PRACTICAL USES

Data anonymization serves various practical purposes across different domains.

1) **Healthcare Research:** In healthcare, anonymized patient data can be shared with researchers while preserving patient privacy [52]. This allows researchers to conduct studies on disease patterns, treatment effectiveness, and public health trends without compromising patient privacy.
2) **Market Research:** Companies often collect large volumes of customer data for market research purposes [53]. Anonymizing this data enables businesses to analyze consumer behavior, preferences, and trends

without accessing personally identifiable information (PII). It helps in understanding market segments and tailoring marketing strategies without infringing on individual privacy.

3) **Data Sharing and Collaboration:** Anonymization facilitates data sharing and collaboration among organizations, particularly in sectors like finance, where sharing transactional data for fraud detection and risk analysis is essential [54]. By anonymizing sensitive financial data, banks, and financial institutions can collaborate with each other or share data with regulatory bodies without disclosing customer identities.
4) **Government and Public Sector:** Governments collect vast amounts of data for statistical analysis, policy-making, and public service planning [12]. Anonymized census data, for example, allows governments to understand demographic trends and allocate resources effectively without compromising citizen privacy. Similarly, anonymizing crime data helps law enforcement agencies analyze crime patterns, formulate crime strategies, and allocate resources for crime prevention without identifying individual suspects.
5) **Academic Research:** Academics often require access to large datasets for research purposes [55]. Anonymizing research datasets allows academic researchers to study various phenomena, such as social behavior, economic trends, and environmental impacts, without exposing personal information about individuals involved in the study.
6) **Machine Learning and Data Analysis:** Anonymized data is used in machine learning and data analysis tasks to develop predictive models, identify patterns, and derive insights without exposing sensitive information [56]. This is particularly important in areas like recommender systems, where user preferences are analyzed to provide personalized recommendations without storing identifiable user data.

D. ADVANTAGES

Data anonymization offers several significant advantages, particularly in the context of privacy protection and data utility.

1) **Enhanced Privacy Protection:** The primary advantage of data anonymization is its ability to protect individual identities [25]. By removing or altering personal identifiers, possible invasions of individual information are significantly reduced. Anonymization helps organizations comply with stringent privacy laws and regulations like GDPR, HIPAA, and CCPA. These laws often require personal data to be anonymized or de-identified before it can be used for secondary purposes.
2) **Data Utilization and Analysis:** Anonymized data can be shared more freely within and between organizations for analysis, research, and other purposes without compromising privacy [57]. Organizations can use anonymized data to gain insights, identify trends, and make informed decisions without the ethical and legal issues associated with using personal data.
3) **Risk Management:** In the event of a data breach, the impact is significantly lessened if the compromised data is anonymized, as it minimizes the risk of direct harm to individuals [58]. By implementing robust anonymization practices, organizations can build trust with customers and stakeholders, enhancing their reputation as responsible data handlers.
4) **Versatility and Flexibility:** Data anonymization is versatile and can be applied in various fields like healthcare, finance, marketing, and public services, making it a universally valuable practice. Anonymization techniques can be adapted to different data types, including structured data like spreadsheets and unstructured data like text and images.
5) **Cost-Effective Data Management:** Anonymizing data can reduce the costs associated with data storage and management, particularly when stringent security measures for personal data are not required. Anonymized data often requires less rigorous handling protocols compared to personal data, thereby streamlining data processing operations.
6) **Innovation and Research:** Anonymized data is invaluable in research settings, allowing for studies that would not be possible with identifiable data due to privacy concerns. Anonymization encourages innovation in data analytics, machine learning, and other technology sectors by providing a safe way to utilize data.

E. LIMITATIONS

While data anonymization offers numerous benefits, it also comes with certain limitations and challenges.

1) **Risk of Re-Identification:** Despite anonymization, sophisticated data mining and machine learning techniques can sometimes re-identify individuals, especially if other data sources are used to cross-reference [59]. Anonymization becomes more challenging with high-dimensional data, as the abundance of attributes increases the risk of re-identification.
2) **Reduction in Data Utility:** Anonymization often involves altering or removing data details, which can reduce the accuracy and granularity of the data, impacting its utility for detailed analysis [60]. For machine learning applications, anonymized data might not provide the level of detail required for accurate model training and predictions.
3) **Complexity and Resource Intensity:** Implementing effective anonymization strategies can be technically complex, requiring specialized knowledge and tools. The process of anonymizing data, especially

large datasets, can be time-consuming and resource-intensive, requiring significant computational power.

4) **Balancing Privacy and Utility:** Finding the right balance between protecting privacy and maintaining data utility is often challenging and context-dependent [61]. As data environments and usage change over time, maintaining this balance requires ongoing assessment and adaptation.
5) **Compliance and Legal Challenges:** Different jurisdictions have varying standards and laws regarding data privacy, making it challenging to ensure compliance, especially for multinational organizations. Inadequate or incorrect anonymization can lead to non-compliance with privacy laws, resulting in legal and financial repercussions.
6) **Ethical and Social Considerations:** There is a risk that anonymized data, especially in sensitive areas like healthcare and finance, can be misused, leading to ethical concerns. Anonymization processes can inadvertently introduce or fail to remove biases present in the data, leading to skewed analyses and decisions.

IV. DATA ENCRYPTION

A. OVERVIEW

Data encryption is a fundamental component of modern cybersecurity, which is crucial in safeguarding information from unauthorized access [62]. This process entails transforming plaintext into ciphertext, an encoded version of the original data, which is nearly impossible to decipher without the correct decryption key. Fig. 4 represents a data encryption and decryption process. It involves an application, an API, and various data storage components. The flowchart shows the transformation of data at different stages: from ''unencrypted app data'' to ''encrypted app data'' and then to ''decrypted data.'' It also highlights the role of ''Vault EaaS'' (Encryption as a Service) in this process. Encryption algorithms, which are mathematical formulas or rule sets, perform this transformation [63]. Well-known algorithms include the Advanced Encryption Standard (AES), Data Encryption Standard (DES), and RSA (Rivest-Shamir- Adleman), each with unique strengths and suited for different applications. Encryption can be broadly classified into two types: symmetric and asymmetric. Symmetric encryption employs the same key for both encryption and decryption, making it fast and efficient for handling large amounts of data [64]. However, the process of distributing the key can be a security risk. On the other hand, asymmetric encryption, which is also called public-key cryptography, requires two keys: one for encryption (the ''public key'') and another (the ''private key'') for decryption. [65]. The method, such as the RSA algorithm, is more secure in key distribution but requires more computational resources. Data encryption has various applications across different fields, such as safeguarding online transactions, emails, and sensitive information stored on computers and servers [66]. Encryption plays a crucial role in the current digital era, where cyber threats and data breaches are common. It acts as a reliable line of defense, ensuring that data remains secure and unintelligible even if they are accessed without authorization or intercepted.

B. TECHNICAL BACKGROUND

1) APPROACH TO DATA ENCRYPTION

The approach to data encryption encompasses several critical components, each tailored to guarantee absolute confidentiality, authenticity, and data integrity. Table 3 compares various data encryption techniques. Advanced Encryption Standard (AES) encrypts data in fixed-size blocks, typically 128 bits; Rivest- Shamir-Adleman (RSA) is based on the concept of asymmetric encryption, which uses two keys: a public key for encryption and a private key for decryption; Homomorphic Encryption allows computations on encrypted data; Proxy Re-encryption, which re-encrypts data without revealing the original plain text; Attribute-Based Encryption, which enables access control based on attributes and Quantum Encryption uses quantum mechanics principles for encryption. The workings of the encryption process are explained below.

1) **Algorithm Selection:** The cornerstone of any encryption strategy is the selection of an appropriate algorithm [67]. This choice hinges on various factors, including the type of data, the level of security required, and the environment in which the encryption is implemented. Symmetric algorithms like AES are preferred for their speed and efficiency in encrypting large data volumes. In contrast, asymmetric algorithms like RSA are more suited for environments where secure key exchange is extremely important.
2) **Key Management and Security:** Effective key management is vital. This involves not only the generation, storage, and destruction of keys but also managing access to keys [68]. Hardware security modules (HSMs) or software-based key management systems may be used in high-security environments to store keys securely.
3) **Security Protocols and Standards:** Aligning with established security protocols and standards is crucial. Protocols like TLS (Transport Layer Security) and standards like FIPS (Federal Information Processing Standards) provide guidelines and best practices for implementing encryption securely [69].
4) **Risk Assessment and Adaptability:** The approach to encryption should include a thorough risk assessment to understand potential threats and vulnerabilities. Encryption strategies must be adaptable to evolving security threats, with regular updates and revisions based on current threat landscapes and advancements in cryptography.
5) **Scalability and Performance:** Data encryption approaches must balance security with performance [70]. Encryption can be resource-intensive. Thus,

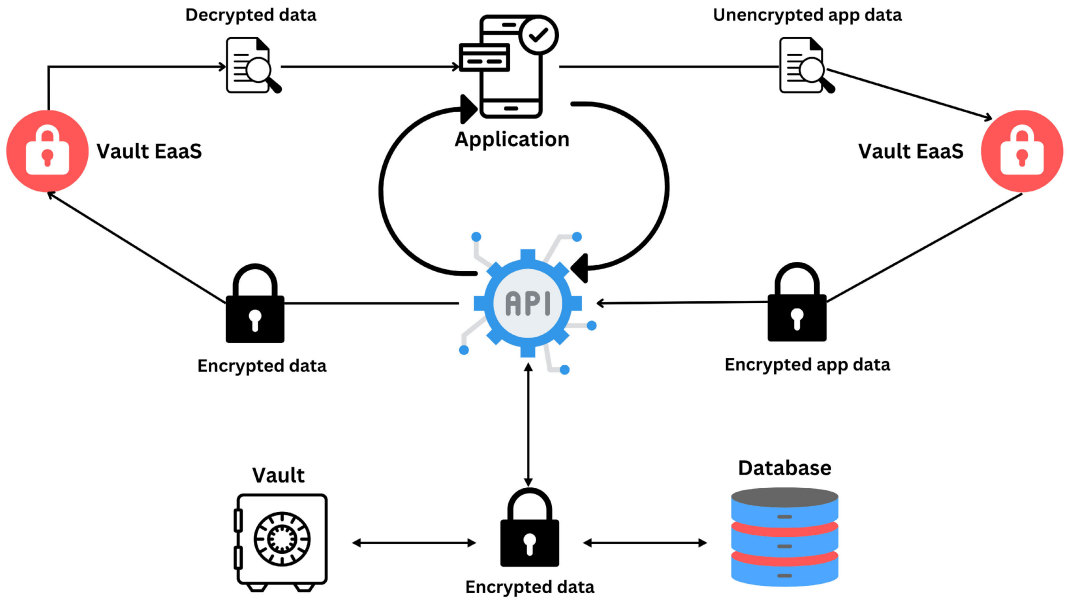

FIGURE 4. Data transmission and encryption across an application.

it is essential to implement a solution that does not significantly degrade system performance and can scale as the organization's data requirements grow.

2) TOOLS & FRAMEWORKS

1) **OpenSSL:** The Secure Sockets Layer (SSL) and Transport Layer Security (TLS) protocols are implemented by OpenSSL, an open-source toolkit, providing robust security solutions for web communication [77]. TLS is a cryptographic protocol that helps secure data transmission over the internet, enhancing SSL with stronger encryption and security. Confidentiality, integrity, and authentication are ensured by encrypting data between client and server. The protocol begins with a handshake to exchange certificates and establish a secure session key for fast symmetric encryption. TLS also uses cryptographic hashes like SHA-256 to maintain data integrity and support session resumption and forward secrecy. Widely implemented through OpenSSL, TLS is crucial for securing online transactions and communications [78].
2) **Crypto++:** Crypto++ is a C++ library of cryptographic algorithms [79]. It includes both symmetric and asymmetric encryption, digital signatures, key exchange protocols, and hash functions like SHA-256. It is known for its high-performance implementations, making it ideal for secure applications that require efficient encryption and decryption, such as data transmission, file security, and communications. As an open-source and cross-platform library, Crypto++ is widely adopted across industries like finance, healthcare, and e-commerce, offering robust and flexible cryptographic solutions to meet diverse security needs.
3) **OpenPGP:** Encrypting and decrypting data with Pretty Good Privacy (PGP) ensures cryptographic privacy and authentication throughout data transmission. [80]. OpenPGP is an open standard that defines the format of encrypted messages. By combining the symmetric and public-key cryptography in a hybrid encryption way, OpenPGP secures the communications. Users create a key pair, consisting of a public key for sharing and a private key for personal use. To encrypt data, the sender generates a session key, encrypts the message with it using symmetric encryption, and secures the session key itself with the recipient's public key. The recipient then decrypts the session key using their private key, which is used to decrypt the message. Key features of OpenPGP include robust encryption, digital signatures for verifying authenticity, tamper detection, and cross-platform compatibility, ensuring secure and reliable data exchange [81].
4) **RSA BSAFE:** RSA BSAFE is a set of security tools and libraries provided by RSA Security [72]. It includes cryptographic libraries that implement various encryption algorithms, key management, and secure random number generation. Its key management features allow secure generation, storage, and distribution of cryptographic keys. RSA BSAFE also incorporates secure random number generation, a critical component for creating encryption keys and ensuring unpredictability in cryptographic operations. The libraries are designed to be highly optimized for performance, supporting various platforms and applications. With a strong focus on compliance, RSA BSAFE aligns with industry standards and regulations, making it a reliable choice for integrating advanced security mechanisms into software systems.

TABLE 3. Comparison of data encryption techniques.

Technique	Description	Advantages	Disadvantages
AES (Advanced Encryption Standard) [71]	AES is a block cipher, meaning it encrypts data in fixed-size blocks, typically 128 bits.	It is highly secure, efficient, and widely adopted for software and hardware encryption.	Key management complexity, potential vulnerabilities if key compromised.
RSA (Rivest-Shamir-Adleman) [72]	RSA is based on asymmetric encryption, which uses two keys: a public key for encryption and a private key for decryption.	Provides strong security for key exchange, digital signatures, and encryption, with widespread use in secure communication protocols.	Slower compared to symmetric encryption, computationally intensive, and requires larger key sizes to ensure strong security.
Homomorphic Encryption [73]	Allowing computations to be performed on encrypted data.	Supports privacy-preserving computation.	Complexity, limited operations supported.
Proxy Re-encryption [74]	Re-encrypting data without revealing the original plain text.	Flexibility in access control preserves privacy.	Key management overhead, computational complexity.
Attribute-Based Encryption [75]	Enabling access control based on attributes instead of identities.	Granular access control supports complex policies.	Key management complexity, computationally intensive.
Quantum Encryption [76]	Leveraging quantum mechanics principles for encryption.	Provides unconditional security, resistant to quantum attacks.	Requires specialized hardware, experimental.

5) **Microsoft.NET Framework:** The .NET Framework includes libraries for implementing data encryption in applications developed using languages such as C# and VB.NET [82]. It supports both symmetric and asymmetric encryption algorithms. The framework provides built-in classes for encryption, including support for AES, RSA, and Triple DES, enabling developers to integrate robust security measures into their applications with minimal effort. Additionally, it facilitates secure data transmission by providing APIs for hashing and digital signatures, ensuring data integrity and confidentiality.
6) **Microsoft SEAL:** Microsoft SEAL (Simple Encrypted Arithmetic Library) is a homomorphic encryption library developed by Microsoft [83]. Homomorphic encryption is a form of encryption that allows computations to be performed on ciphertexts, producing encrypted results that, when decrypted, match the outcome of operations performed on the plaintext. This is particularly useful for privacy-preserving computations in scenarios like cloud computing, where data can remain encrypted while being processed. Microsoft SEAL supports both CKKS and BFV encryption schemes, enabling efficient computation on encrypted integers and real numbers. Its modular design allows seamless integration into various programming environments, making it a versatile tool for privacy-preserving AI and secure data analysis.
7) **AWS Key Management Service (KMS):** AWS KMS is a fully managed key management service provided by Amazon Web Services [84]. It allows users to create and control encryption keys used to encrypt their data. AWS KMS integrates with other AWS services like S3, DynamoDB, and Lambda, enabling secure key management across cloud applications. It also supports features like automated key rotation and detailed audit logging, ensuring that keys are not only secure but also compliant with regulatory standards.

3) EVALUATING ENCRYPTION EFFICACY

1) **Security Efficacy [85]:**
   a) Criteria: Conduct regular red teaming for testing the security of systems with simulated attacks.
   b) Evaluation: Analyze encryption algorithms' resilience to advanced threats, ensuring a comprehensive assessment of security effectiveness.
2) **Key Management [68]:**
   a) Criteria: Implement hardware security modules for key protection.
   b) Evaluation: Regularly audit key generation, storage, and distribution processes, addressing vulnerabilities in the key management life cycle.
3) **Performance Impact [86]:**
   a) Criteria: Employ benchmarking tools for performance analysis.
   b) Evaluation: Test encryption under varying workloads to identify resource-intensive operations and optimize algorithms.
4) **Compliance [26]:**
   a) Criteria: Establish a compliance framework.
   b) Evaluation: Periodically review and update encryption protocols to align with evolving data protection regulations, avoiding legal and regulatory penalties.
5) **Usability and Transparency [87]:**
   a) Criteria: Implement user-friendly encryption interfaces.
   b) Evaluation: Collect user feedback on encryption processes, iteratively improving interfaces to enhance user understanding and acceptance.
6) **Scalability [70]:**
   a) Criteria: Design encryption solutions for horizontal scalability.
   b) Evaluation: Perform stress tests to ensure the encryption system can handle increased data volumes without compromising performance.

7) **Interoperability [88]:**
   a) Criteria: Develop encryption solutions compatible with common standards.
   b) Evaluation: Test integration with existing IT infrastructure components to guarantee smooth interoperability.
8) **Resilience Against Emerging Threats [24]:**
   a) Criteria: Establish threat intelligence integration.
   b) Evaluation: Regularly update encryption protocols based on the latest threat intelligence to stay ahead of emerging risks.

C. PRACTICAL USES

As digital technologies advance and cyber threats become more sophisticated, ensuring sensitive data's confidentiality, integrity, and authenticity has become most important. The remaining subsection describes practical applications of data encryption across various domains, highlighting its significance in safeguarding information during transit, storage, and data processing.

1) **Securing Communication Channels:** Encryption is essential for protecting sensitive information transmitted over networks [89]. Secure Sockets Layer (SSL) and Transport Layer Security (TLS) protocols employ encryption to secure communication channels, preventing eavesdropping and data tampering.
2) **E-commerce and Online Transactions:** Data encryption is critical in e-commerce platforms to secure online transactions [90]. Secure encryption algorithms, such as AES, protect financial information and personal details during payment processes, instilling trust in online consumers.
3) **Healthcare Data Protection:** In the healthcare sector, where patient confidentiality is essential, data encryption ensures the protection of electronic health records (EHRs) and sensitive medical information [91]. Compliance with regulations like the Health Insurance Portability and Accountability Act (HIPAA) mandates the use of encryption to safeguard patient data.
4) **Cloud Computing Security:** Cloud services rely heavily on data encryption to maintain the privacy and security of stored data [92]. Client-side encryption, secure key management, and end-to-end encryption are deployed to safeguard data in transit and at rest, even when stored in third-party cloud providers.
5) **Financial Institutions and Banking:** Banks and financial institutions utilize encryption to secure financial transactions, customer account information, and sensitive financial data [93]. Encryption helps prevent fraud, unauthorized access, and data breaches in the financial sector.
6) **Government and National Security:** Governments employ data encryption to protect classified information, secure communications, and ensure the integrity of sensitive data related to national security [94]. Encryption is an integral part of securing government databases, communication networks, and critical infrastructure.

D. ADVANTAGES

1) **Confidentiality:** Encryption ensures that only authorized individuals or systems can access the encrypted data [95]. It prevents unauthorized users from understanding or interpreting the content, maintaining the confidentiality of sensitive information.
2) **Data Integrity:** Encryption helps maintain data integrity by detecting and preventing any unauthorized alterations [62]. Cryptographic algorithms generate unique checksums or hashes, enabling users to check for data manipulation during storage or transfer.
3) **Protection Against Unauthorized Access:** Encrypted data is significantly more resistant to unauthorized access. Even if a malicious actor gains access to the encrypted information, without the proper decryption key, the data remains unintelligible and inaccessible.
4) **Secure Communication:** In the context of communication over networks, encryption ensures the security of transmitted data [89]. Protocols like SSL/TLS use encryption to protect data exchanged between users and servers, preventing eavesdropping and man-in-the-middle attacks.
5) **Secure Storage in the Cloud:** Cloud service providers use encryption to secure data stored in cloud environments. Client-side encryption and other techniques ensure that even if there is a breach in the cloud infrastructure, the data remains protected from unauthorized access.
6) **Protection Against Ransomware:** Encryption provides a defense mechanism against ransomware attacks. If files are encrypted before being targeted by ransomware, the attacker is unable to access the data without the decryption key, thwarting their attempts to extort the user or organization.

E. LIMITATIONS

While data encryption is a powerful and widely used security measure, it certainly comes with certain limitations and challenges.

1) **Key Management Complexity:** Managing encryption keys can be complex, especially in large-scale systems [96]. The secure generation, distribution, storage, and rotation of encryption keys require careful planning. If keys are lost or compromised, it can lead to data loss or unauthorized access.
2) **Performance Overhead:** Encryption and decryption processes can introduce a computational overhead, slowing down data processing and transmission. This performance impact may be more noticeable in resource-constrained environments, such as IoT devices or systems with limited computing power.

3) **Data Accessibility:** While encryption protects data from unauthorized access, it also introduces challenges related to data accessibility. Authorized users must have access to the decryption keys, which can be a hurdle in emergency situations or when key holders are unavailable.
4) **Insider Threats:** Encryption does not protect against insider threats where individuals with legitimate access misuse their privileges. Authorized users with access to encryption keys can potentially abuse their permissions to compromise sensitive data.
5) **Side-Channel Attacks:** Side-channel attacks exploit leaked information during encryption or decryption, such as timing patterns, power consumption, or electromagnetic emissions. These attacks can reveal encryption keys and compromise system security.
6) **Quantum Computing Threats:** Quantum computing is dangerous to current encryption techniques [76]. Quantum computers have the potential to defeat popular encryption systems, highlighting the need for quantum-resistant encryption algorithms.
7) **Costs and Resource Requirements:** Implementing robust encryption solutions can incur costs, both in terms of hardware and software. Resource-intensive encryption processes may require additional computational power, increasing infrastructure costs.

V. SYNTHETIC DATA

A. OVERVIEW

In the dynamic landscape of contemporary data science and artificial intelligence, the criticality of access to high-quality, diverse datasets for model development and training cannot be overstated [97]. The advent of synthetic data represents a paradigm shift in addressing persistent challenges surrounding data availability, privacy concerns, and the need for robust machine learning models. Synthetic data refers to artificially generated datasets that emulate the statistical properties of real-world data without containing any sensitive or personally identifiable information [98]. This innovative approach has gained traction across various industries as it provides a means to navigate through the delicate balance between data utility and individual privacy, thereby facilitating advancements in machine learning, data analytics, and artificial intelligence applications.

Synthetic data finds applications across diverse domains, including healthcare, finance, cybersecurity, and beyond. In healthcare, where access to large-scale, diverse patient data is often restricted due to privacy concerns, synthetic data allows researchers to train and validate models without compromising sensitive information. In financial sectors, synthetic data proves invaluable for stress testing models and evaluating system robustness without exposing real financial data. Furthermore, in cybersecurity, synthetic datasets serve as a crucial tool for training and testing intrusion detection systems, enabling the development of more resilient defenses against evolving cyber threats. Fig. 5 is a flowchart related to data synthesis and privacy assurance services. It outlines a process that starts with real data, goes through synthesis to produce synthetic data, and includes steps for utility assessment, privacy assurance, and report generation.

B. TECHNICAL BACKGROUND

1) APPROACH TO SYNTHETIC DATA

This section delves into the diverse landscape of synthetic data generation techniques, providing a comprehensive overview of the state-of-the-art methods employed in addressing data scarcity and privacy concerns across various domains. The synthesis of realistic and privacy-preserving data has become a critical facet of contemporary research, as it enables the development and evaluation of robust machine-learning models. Table 4 compares various synthetic data generation techniques. It includes Generative Adversarial Networks (GANs), which use two neural networks to generate realistic data; the Copulas-based method generates synthetic data by capturing the dependencies between multiple variables, especially in multivariate distributions; Synthetic Minority Over-sampling Technique (SMOTE) generates new, synthetic samples by interpolating between existing minority class instances; the transformer-based model uses large original datasets to understand the structure and typical distribution of data and Variational Autoencoders (VAEs), which learn latent representations to generate data. The workings of synthetic data generation are explained below.

1) **Synthesis Technique Selection:** The foundation of the synthetic data generation process lies in the selection of an appropriate synthesis technique [104]. Choosing the right synthesis technique is critical in synthetic data generation, as it determines the quality, realism, and utility of the output. The choice depends on dataset characteristics, required realism, and application goals. Techniques such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and rule-based methods each offer distinct advantages. GANs, involving a generator and discriminator in an adversarial setup, produce highly realistic data, making them ideal for applications like image generation and time-series analysis, though they can be computationally demanding and unstable. VAEs, which encode and decode data through a latent space, excel in creating diverse outputs for tasks like healthcare monitoring but may produce less sharp results than GANs. Rule-based methods, which use predefined logic to generate structured data, are particularly suited for fields like finance and energy, where consistency is essential. Other methods, such as SMOTE, Data Synthesizer (DS), and SynthPop Non-Parametric (SP-NP), address specific challenges. SMOTE generates data for minority classes to tackle imbalances, while DS ensures privacy through differential privacy techniques. SP-NP

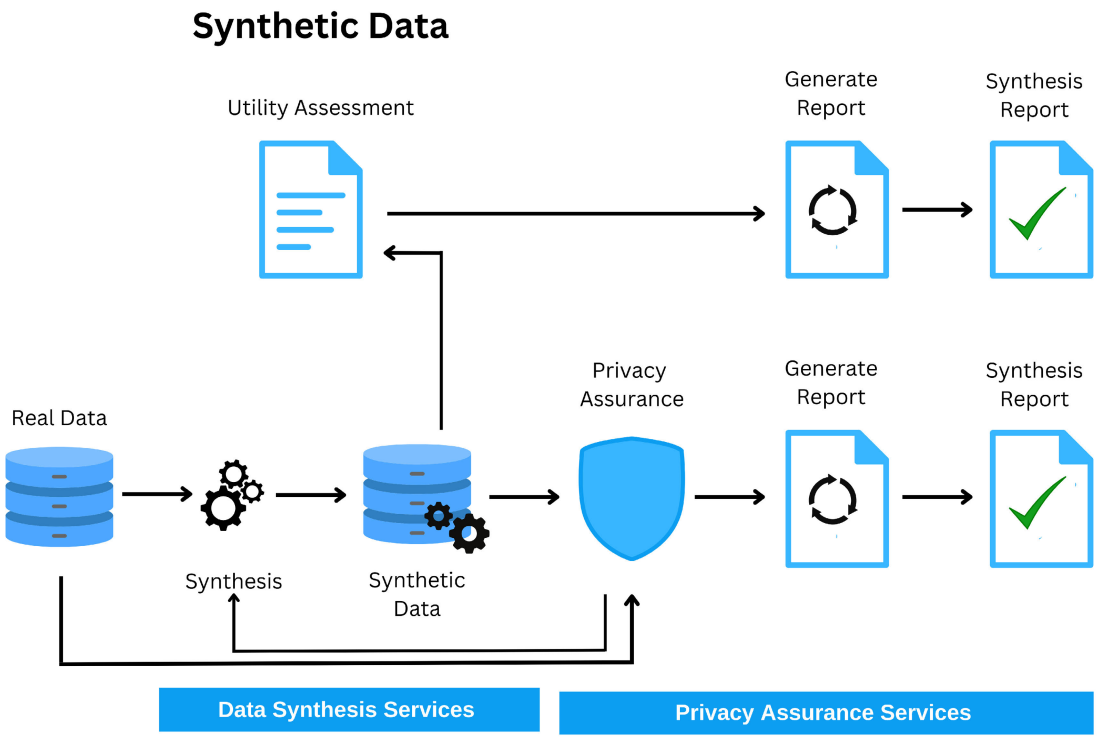

FIGURE 5. Synthetic data analyzation.

TABLE 4. Comparison of synthetic data generation techniques.

Technique	Description	Advantages	Disadvantages
Generative Adversarial Networks (GANs) [99]	Two neural networks compete to generate realistic data.	Preserves statistical properties and maintains privacy.	Training complexity, the potential for mode collapse.
Copulas-based method [100]	Generates synthetic data by capturing the dependencies between multiple variables, especially in multivariate distributions.	Copulas enables the generation of synthetic data that maintains realistic relationships between variables, preserving individual behavior (marginals) and joint behavior (dependencies).	It is computationally complex in high dimensions and is sensitive to outliers and wrong dependence assumptions.
SMOTE (Synthetic Minority Over-sampling Technique) [101]	SMOTE generates new, synthetic samples by interpolating between existing minority class instances.	It is easy to use and improves model performance on imbalanced data by generating diverse minority class samples.	It may create overlapping samples and struggle with high-dimensional or mixed-type data.
Transformer-based model [102]	GPT-based models use large original datasets to understand the structure and typical distribution of data.	Transformers capture long-range dependencies, are highly scalable, versatile, and handle complex patterns well.	They are computationally expensive and require a huge amount of data.
Variational Autoencoders (VAEs) [103]	Learning latent representations to generate data.	Captures data distribution provides control over a generation.	Requires large datasets for training and complex optimization.

preserves probabilistic relationships within data, making it useful in healthcare and engineering. Ultimately, the technique should align with the dataset's needs and the application's objectives, balancing factors like computational efficiency, stability, and data quality. [105]

2) **Synthetic Data and Anonymization:** The hybrid technique combines anonymization and synthetic data generation using a permutation paradigm. Synthetic data are created based on attribute ranks, avoiding direct mapping to the original data. k-Anonymity is then applied to ensure each synthetic record is indistinguishable from at least k others, mitigating reidentification risks. This approach preserves attribute relationships while balancing data utility and strong privacy guarantees, making it ideal for secure data sharing [106].
3) **Feature Representation and Modeling:** Feature representation and modeling are crucial in ensuring that synthetic data retains the key characteristics and relationships of the original dataset. Statistical analyses provide insights into data structure, while correlation

assessments help preserve dependencies between variables. Feature importance analysis further ensures that critical attributes are accurately represented. Deep learning models, such as convolutional networks, are commonly used to extract hierarchical features, capturing both basic and complex patterns. Techniques like gating mechanisms and dropout layers enhance generalization and reduce overfitting, particularly in scenarios with limited data. [107]. These approaches collectively enable the creation of synthetic datasets that reflect real-world patterns and maintain their utility for various applications.

4) **Privacy-Preserving Measures:** Privacy considerations are integrated into the synthetic data generation approach [108]. Techniques such as differential privacy, homomorphic encryption, and data perturbation are employed to safeguard individual privacy while maintaining the utility of the synthetic data. The level of privacy preservation is adapted based on the sensitivity of the data and regulatory requirements.
5) **Evaluation Metrics for Synthetic Data Quality:** Rigorous evaluation metrics are employed to assess the quality of the synthetic data [109]. Distributional similarity measures, task-specific performance metrics, and domain-specific evaluation criteria are utilized to ensure that the synthetic data aligns with the intended use cases and effectively represents the underlying data distribution.
6) **Adaptability to Evolving Data Landscapes:** A dynamic approach to synthetic data generation involves regular risk assessments to detect potential threats and vulnerabilities [110]. The synthesis strategy must be adaptable to evolving security and privacy challenges, with provisions for updates and revisions based on advancements in generative techniques and changes in the data landscape.
7) **Computational Efficiency and Scalability:** Balancing computational efficiency with synthesis quality is imperative [111]. The synthetic data generation approach is optimized for scalability, employing parallelization techniques and efficient algorithms. This ensures that the synthetic data generation process remains feasible for large-scale datasets without significantly compromising computational performance.

2) TOOLS & FRAMEWORKS

1) **Synthetic Data Generation Libraries:**
   a) Faker: Faker is a Python library that generates realistic but non-sensitive data [112]. It can be used to create synthetic datasets with names, addresses, dates, and other attributes that mimic real-world data. Faker is highly customizable, allowing users to define schemes for generating data specific to their needs. Its integration with Python-based data pipelines makes it an essential tool for testing systems, simulating real-world scenarios, and preserving data privacy in various applications [112]. Additionally, Faker supports localization, enabling the generation of data tailored to specific cultural or regional contexts [112].
   b) Mimesis: Similar to Faker, Mimesis is another Python library for data generation [113]. It supports a wide range of data types and locales, making it versatile for creating synthetic datasets for various purposes. Mimesis is particularly efficient in generating large-scale synthetic workloads and datasets [113]. Its advanced features include support for metadata workload modeling, which allows it to simulate complex storage and data management systems. This capability makes Mimesis an excellent tool for testing distributed systems, evaluating performance benchmarks, and conducting stress tests under realistic workload conditions [113].
2) **Deep Learning Frameworks:**
   a) TensorFlow: An open-source deep learning framework widely used for building and training neural networks. TensorFlow provides tools for creating and training generative models like GANs and VAEs, making it a popular choice for synthetic data generation in machine learning applications. TensorFlow also supports distributed training, enabling large-scale model development across multiple machines. Its compatibility with TensorFlow Privacy further enhances its utility for privacy-preserving generative model training [7], [114].
   b) PyTorch: Another popular deep learning framework, PyTorch is known for its dynamic computation graph and is often used for developing generative models. It is praised for its simplicity and flexibility, making it suitable for various research and development tasks. PyTorch also includes robust debugging capabilities and visualization tools like TensorBoard, simplifying the process of understanding and optimizing model performance. Its active community and extensive library support make it ideal for experimenting with advanced deep learning techniques [109], [114].
3) **Generative Models:**
   a) GANs (Generative Adversarial Networks): GANs consist of a generator and a discriminator network that is trained adversarially. TensorFlow and PyTorch provide implementations and pre-trained models for GANs, allowing users to generate synthetic data by learning from the patterns of the original dataset. Recent advancements in GANs, such as Differentially

Private GANs (DP-GANs), enable the generation of synthetic data with strong privacy guarantees by incorporating differential privacy techniques. These developments are particularly useful for privacy-sensitive applications in domains like healthcare and finance [114], [115].

b) VAEs (Variational Autoencoders): VAEs are generative models that learn a probabilistic mapping between the input and a latent space. TensorFlow and PyTorch support VAE implementations, enabling the generation of synthetic data by sampling from the learned latent space. Incorporating privacy-preserving mechanisms into VAEs has led to the development of Differentially Private VAEs, which balance data utility and privacy, making them suitable for applications in sensitive areas like IoT and personalized AI [114], [116].

4) **Data Augmentation Tools:**

a) imgaug: A powerful Python library for augmenting images [117]. It offers a variety of transformations such as rotation, scaling, and flipping, which can be useful when working with image datasets.

b) Albumentations: Another image augmentation library that is designed to be fast and efficient [118]. It supports a wide range of augmentation techniques and is particularly popular in computer vision applications.

5) **Privacy-Preserving Tools:**

a) PySyft: PySyft is a Python library for encrypted, privacy-preserving machine learning. It supports techniques such as Federated Learning and Homomorphic Encryption, enabling the generation of synthetic data with privacy considerations.

b) PyDifferential-Privacy: A library focused on implementing differential privacy in Python. It provides tools for adding controlled noise to data and preserving individual privacy during synthetic data generation.

6) **Rule-Based Synthetic Data Generation:**

a) SDV (Synthetic Data Vault): SDV is a Python library that employs rule-based methods for synthetic data generation. It allows users to define rules and constraints to generate synthetic datasets that adhere to specific criteria. SDV also includes support for relational data generation, enabling the synthesis of multi-table datasets while maintaining relationships and dependencies across tables. This makes it particularly useful for applications in database testing and data privacy research [109], [114].

3) EVALUATING SYNTHETIC DATA EFFICACY

1) **Distributional Similarity:** This involves assessing how closely the statistical distribution of synthetic data matches that of the original dataset. Metrics like Kolmogorov-Smirnov or Wasserstein distance quantify the divergence between the distributions. It ensures that the synthetic data captures the essential statistical characteristics of the real data, supporting meaningful analysis and model training.

2) **Task-Specific Model Performance:** In this criterion, the focus is on the performance of machine learning models trained on synthetic data. The success of synthetic data is determined by how well models generalize to real-world scenarios. Tasks may include classification, regression, or clustering. If models trained on synthetic data exhibit performance comparable to those trained on real data, it signifies the utility and effectiveness of the synthetic dataset for specific applications.

3) **Privacy Preservation:** Privacy is a critical consideration. Evaluation involves assessing the impact of privacy-preserving techniques implemented during synthetic data generation. Differential privacy, homomorphic encryption, or other methods should be effective in preventing the identification of individuals in the synthetic dataset while preserving the overall statistical properties.

4) **Realism and Diversity:** Realism refers to how well the synthetic data captures the patterns and relationships present in the original dataset. Diversity assesses the variability of the synthetic data. Both are crucial for ensuring that synthetic datasets are representative of real-world scenarios, enhancing the applicability of generated data in diverse use cases.

5) **Generalization:** Generalization measures how well models trained on synthetic data perform on unseen, real-world data. The synthetic dataset should not only capture the characteristics of the training data but also enable models to adapt effectively to new, unseen data. Robust generalization is vital for the practical applicability of synthetic data in different contexts.

6) **Scalability:** Scalability assesses the efficiency of the synthetic data generation process as the dataset size increases. Successful synthetic data approaches should be capable of generating realistic samples for large datasets without a substantial increase in computational demands. Scalability is crucial for practical applications where data volumes are significant.

7) **Ethical Considerations:** Ethical considerations involve evaluating the fairness and bias in synthetic data. The synthetic dataset should reflect the diversity present in the original data and avoid introducing or perpetuating biases. Assessments should involve qualitative analysis and expert judgment to ensure the ethical use of synthetic data in various applications.

C. PRACTICAL USES

1) **Healthcare Applications:** Synthetic data has found significant applications in the healthcare industry.

Various healthcare institutions are leveraging synthetic data to model and conduct tests in scenarios where real data is not available [119]. For example, the U.S. National Institutes of Health is collaborating with Syntegra, an IT services firm, to create a de-identified version of their COVID-19 patient records database. This synthetic dataset can be shared with researchers worldwide, facilitating global research efforts to understand the disease and expedite the development of treatments and vaccines.

2) **Autonomous Vehicles:** The autonomous vehicle industry is witnessing a revolution with the advent of synthetic data. Pioneers in this field, such as Waymo and Tesla, are harnessing the power of synthetic data to train their self-driving algorithms [120]. This artificial data enables the simulation of diverse driving conditions and scenarios, which is indispensable for the evolution and testing of autonomous vehicles. By merging techniques from the film and gaming industries (like simulation and CGI) with generative neural networks (such as GANs and VAEs), car manufacturers can create realistic datasets and simulated environments at scale without the need for real-world driving. Synthetic data is accelerating the training programs of autonomous vehicle developers by allowing their algorithms to test their capabilities round-the-clock in a purely digital environment.
3) **Retail Industry:** The retail sector is exploring new revenue opportunities with synthetic data. Retailers can monetize synthetic data reflecting customer purchasing behavior without compromising personal information. Companies like H&M have already started using bots to gather user preferences and tailor their advertising campaigns. Synthetic data allows the generation of large volumes of realistic data without the need to collect or process extensive real-world data [121]. This enhances privacy as there's no need to expose or share personally identifiable information (PII). Synthetic data has emerged as a solution that ensures data quality for training and validating machine learning models in retail. It enables retailers to create realistic yet artificial datasets that replicate the characteristics of real-world data.
4) **Finance Sector:** Synthetic data is becoming a valuable tool in the financial industry. It allows for the simulation of financial markets, enabling the testing of trading strategies and risk models without the need for real market data. Furthermore, synthetic data is being utilized to identify and mitigate bias in customer interactions while adhering to data privacy laws [122]. Financial institutions generate a vast amount of complex and diverse data. However, sharing this data within different business units and outside the organization is often restricted due to regulatory requirements and business needs. Synthetic data offers a solution to this problem, producing financial datasets with real-world qualities while maintaining privacy for all parties involved.

D. ADVANTAGES

Synthetic data presents numerous advantages across diverse domains:

1) **Privacy Preservation:** Synthetic data generation allows for the creation of datasets that do not contain sensitive or personally identifiable information, ensuring privacy and compliance with data protection regulations [12]. Synthetic data can be used to share insights or test algorithms without revealing the details of the original dataset, protecting individuals' privacy.
2) **Data Diversity:** Synthetic data generation methods enable the creation of datasets that capture the statistical characteristics of real-world data [109]. This diversity is valuable for training machine learning models on a wide range of scenarios, enhancing their generalization capabilities. Synthetic data can be designed to include rare or extreme events that may be insufficiently represented in real-world datasets, allowing for more robust model testing.
3) **Data Augmentation:** In scenarios where obtaining large real-world datasets is challenging, synthetic data serves as a useful tool for augmenting the existing data [123]. This aids in preventing overfitting and improving the performance of machine learning models. Synthetic data generation allows researchers to manipulate and engineer features, creating datasets that emphasize specific aspects of the data distribution.
4) **Reduced Bias and Fairness Testing:** Synthetic data can be generated to reduce biases present in real-world datasets, fostering fairness and reducing the risk of biased algorithmic decisions [124]. Synthetic data facilitates the testing of models for fairness by introducing controlled variations and assessing their impact on model predictions.
5) **Security and Testing:** Synthetic data provides a safe environment for testing security protocols, algorithms, and models without exposing real-world vulnerabilities [125]. Synthetic data can be used to assess the robustness of systems and algorithms by simulating diverse scenarios, including potential adversarial attacks.
6) **Cost-Effective Solution:** Generating synthetic data can be more cost-effective and efficient compared to collecting and managing large volumes of real-world data, especially in situations where acquiring real data is time-consuming or expensive.
7) **Flexibility in Experimentation:** Researchers can simulate specific scenarios or edge cases with synthetic data, allowing for controlled experimentation and analysis. Synthetic data facilitates iterative testing and refinement of algorithms, making it easier to identify and address issues in models before deploying them in real-world settings.

8) **Data Sharing and Collaboration:** Synthetic data enables collaborative research without the need to share sensitive or proprietary real-world datasets, promoting open science and fostering research collaboration. By preserving the statistical properties of the original data while ensuring privacy, synthetic data enables secure sharing across institutions. This is particularly valuable in fields like healthcare and finance, where regulatory constraints limit real data sharing. Additionally, tools like SDV and TensorFlow Privacy support the generation of privacy-compliant synthetic data for collaborative analytics [25], [114].

E. LIMITATIONS

While synthetic data offers various advantages, it comes with certain limitations that researchers and practitioners need to consider:

1) **Privacy Risks in Incomplete Anonymization:** Incomplete anonymization of synthetic data poses privacy risks, as sophisticated attackers might still be able to identify individuals or sensitive information, undermining the privacy benefits of synthetic datasets.
2) **Difficulty in Modeling Complex Interactions:** Capturing intricate relationships and complex interactions within the data is challenging for synthetic data generation methods. As a result, models trained on synthetic data may not generalize well to real-world complexities.
3) **Lack of Real-World Variability:** Synthetic data may not fully capture the complexity and variability present in real-world datasets. This limitation can impact the performance of models when applied to diverse and unpredictable scenarios.
4) **Dependency on Accurate Modeling Assumptions:** The effectiveness of synthetic data generation relies on accurate modeling assumptions. Deviations from these assumptions may lead to synthetic datasets that do not accurately reflect the characteristics of the real-world data.
5) **Challenge in Handling Domain-Specific Knowledge:** Incorporating domain-specific knowledge into synthetic data generation is challenging. Certain fields may have unique data characteristics that are difficult to replicate without an in-depth understanding of the domain.
6) **Complexity in Dynamic Environments:** Adapting synthetic data generation methods to dynamic environments, where data distributions change over time, is complex. Keeping synthetic datasets up-to-date and reflective of evolving real-world conditions can be a significant challenge.

VI. DIFFERENTIAL PRIVACY

A. OVERVIEW

In an age where vast amounts of data are generated and shared, maintaining individual privacy has become a crucial challenge. Differential privacy (DP) has emerged as a powerful framework for protecting sensitive information in datasets while still allowing for meaningful data analysis. This technique seeks to provide a quantifiable privacy guarantee by ensuring that the removal or addition of a single data point does not significantly affect the outcome of any analysis performed on the dataset [126].

The primary objective of differential privacy is to protect individuals from re-identification and privacy breaches, even when their data is included in a dataset used for analysis. Unlike traditional anonymization techniques, which focus on removing identifiable information, differential privacy mathematically limits the amount of information an observer can infer about any individual. This is achieved by adding a small amount of controlled noise to the data or the results of computations, which obscures the specific contributions of individual data points without significantly diminishing the overall utility of the dataset [127].

Differential privacy is widely applicable in various domains, from healthcare to financial services, where sensitive personal data is handled regularly. Its adoption is particularly important as organizations navigate the complexities of data privacy regulations like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). As data breaches and misuse of personal data become increasingly frequent, the need for robust privacy-preserving mechanisms such as differential privacy has never been more critical [126]. At its core, differential privacy represents a shift from simply removing identifiers to focusing on statistical methods that ensure privacy at the mathematical level. This approach provides strong privacy guarantees while enabling organizations to leverage valuable insights from large-scale datasets without compromising individuals' privacy. As the demand for privacy-preserving data analytics continues to grow, differential privacy stands out as a key solution for the future of data security.

B. TECHNICAL BACKGROUND

The primary goal of differential privacy is to modify datasets to protect individual privacy while preserving the data's utility for meaningful analysis and decision-making. Unlike traditional anonymization techniques, differential privacy provides a mathematically rigorous framework that quantifies privacy guarantees even against adversaries with significant background knowledge. Table 5 compares various differential privacy techniques. Below, we have discussed the foundational principles, mechanisms, and advancements in differential privacy.

1) Noise Mechanism in Differential Privacy: The noise mechanism is central to differential privacy, protecting sensitive information while preserving data utility. By adding controlled random noise to query outputs or datasets, it masks the presence or absence of any single individual's data. The amount and type of noise depend on the query's sensitivity, the

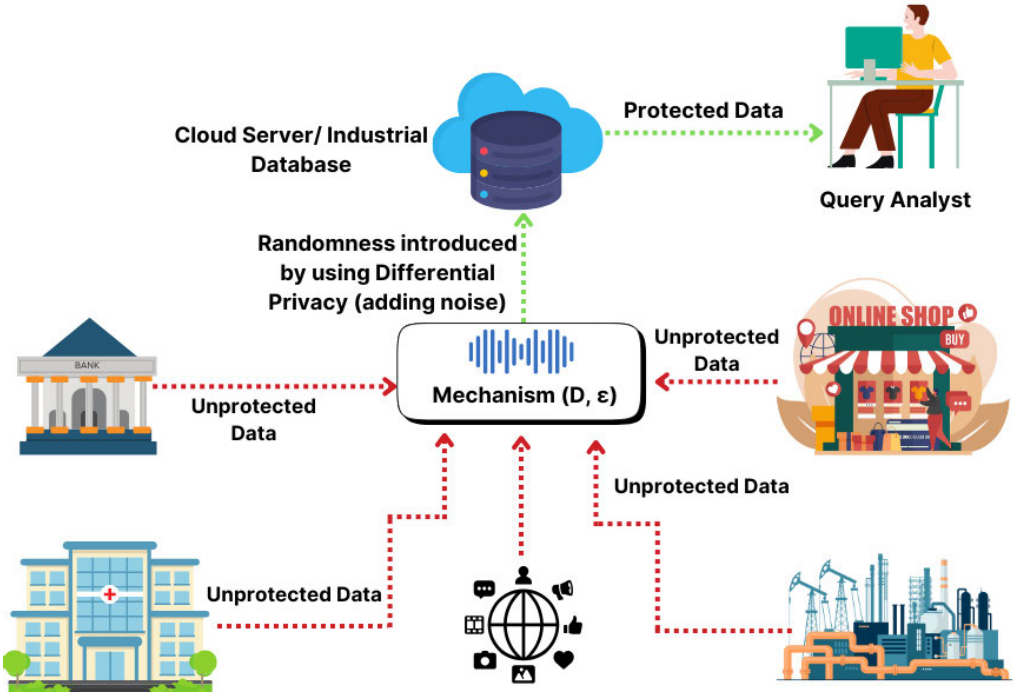

FIGURE 6. Differential Privacy across different applications.

privacy budget (ϵ) balancing privacy and accuracy, and the data type being processed. Carefully calibrated noise ensures outputs remain statistically close to the original values, enabling meaningful analysis while maintaining privacy. The Laplace mechanism adds noise from a Laplace distribution, proportional to the query sensitivity and inversely proportional to (ϵ) [128]. It is effective for numeric queries like sums and averages. The Gaussian mechanism introduces noise from a Gaussian distribution and is widely used in high-dimensional datasets and machine learning, particularly for approximate differential privacy [7]. These mechanisms collectively provide robust protection while enabling diverse analytical and machine-learning applications.

2) Variants of Differential Privacy: Local Differential Privacy (LDP) focuses on protecting privacy at the data source by allowing individuals to perturb their data before sharing it with a data collector. This eliminates the need for a trusted central aggregator, making it particularly suitable for distributed systems. LDP is highly effective in decentralized environments where privacy must be maintained even without centralized trust, ensuring robust protection for sensitive data [128]. Another variant, Rényi Differential Privacy, extends the standard differential privacy framework by incorporating different privacy loss distributions. This approach provides greater flexibility in balancing privacy and utility, particularly in scenarios involving iterative computations or high-complexity systems. Rényi Differential Privacy is increasingly recognized for its ability to adapt to advanced computational frameworks while maintaining robust privacy guarantees [126]. These variants enable tailored privacy solutions across diverse applications, from decentralized networks to cutting-edge quantum technologies.

3) Privacy Budget and Trade-offs: The privacy budget (ϵ) is a fundamental parameter in differential privacy, determining the amount of noise added to data and the balance between privacy protection and utility. Smaller values of ϵ provide stronger privacy guarantees by introducing more noise, but they can significantly reduce the accuracy of the data. Conversely, larger ϵ values improve data utility but increase the risk of privacy breaches. Selecting an optimal ϵ value requires carefully considering the specific application and its privacy-utility trade-offs. In scenarios like machine learning, where maintaining model accuracy is critical, tuning ϵ involves iterative adjustments and sensitivity analysis to achieve a suitable balance. This process often includes privacy accounting techniques to track cumulative privacy loss over multiple queries, ensuring robust protection while preserving the dataset's analytical value. In advanced computational systems like quantum frameworks, where computations may involve iterative algorithms, calibrating ϵ becomes even more crucial to maintain privacy and computational efficiency [126], [129].

1) APPROACH TO DIFFERENTIAL PRIVACY

Implementing differential privacy requires a structured and thoughtful process to protect privacy while maintaining data

utility. The methodology involves several critical steps, from identifying the context and risks to selecting and applying the appropriate mechanisms. Fig 6 illustrates how differential privacy can be applied to different real-world applications. The following are the key steps and considerations:

1) **Understanding the Privacy Context:**
   a) Identifying Sensitive Data: The first step in applying differential privacy is understanding which data elements are sensitive and require protection. Datasets, particularly those involving interconnected relationships, often contain elements like social connections or attributes that can reveal identities if not adequately protected [130].
   b) Defining the Use Case: The purpose for which the data will be used significantly affects how differential privacy is implemented. For instance, public data releases, research studies, and machine learning models have different privacy requirements [7].
2) **Risk Assessment and Budgeting:**
   a) Before applying differential privacy, it is essential to identify potential risks, such as re-identification attacks or adversarial exploitation. Advanced computational environments, such as quantum systems, pose unique challenges where adversaries may employ sophisticated attack scenarios. Proper modeling of these risks ensures that differential privacy mechanisms are robust and capable of protecting sensitive data against potential threats [126].
   b) Determining the Privacy Budget (ϵ): The privacy budget (ϵ) is a critical parameter that defines the trade-off between privacy protection and data utility. Smaller values of ϵ provide stronger privacy but may reduce the data's practical usefulness. Selecting an optimal ϵ value requires careful consideration of specific use cases and acceptable trade-offs. This involves understanding the sensitivity of the data and how much noise can be introduced without overly compromising the accuracy of the analysis [129].
3) **Selecting the Appropriate Mechanism:**
   a) Noise Addition Techniques: The choice of noise mechanism depends on the type of data and the privacy model being used. For numeric data, the Laplace mechanism is commonly employed, introducing noise proportional to the sensitivity of the query and inversely proportional to the privacy budget (ϵ). This method is particularly effective for aggregate queries like sums and averages, ensuring individual data points remain indistinguishable. For approximate privacy or high-dimensional data, the Gaussian mechanism is often used due to its flexibility and its ability to balance privacy and utility in scenarios where smaller noise may suffice [131].
   b) Local vs. Central Differential Privacy: The implementation approach also depends on whether a centralized or decentralized system is used. Local Differential Privacy (LDP) is particularly effective in distributed environments, where individuals perturb their data locally before sharing it with a data collector. This ensures privacy at the source and is especially valuable when a trusted central aggregator cannot be assumed. Central differential privacy, on the other hand, applies noise at the aggregator level, making it more suitable for centralized data collection systems [128].
   c) Differential Privacy and Synthetic Data: Generating synthetic data with differential privacy (DP) involves creating a dataset that mimics the statistical properties of real data while ensuring individual privacy through formal DP guarantees. This process typically adds noise to statistical queries or during model training to prevent sensitive information from being inferred. Generative models like DP-GAN [132] and DP-CTGAN [133] can produce synthetic data by incorporating DP mechanisms such as DP-SGD (Differentially Private Stochastic Gradient Descent), where gradients are clipped, and noise is added to preserve privacy. The privacy budget (ϵ) is carefully managed to balance privacy and data utility, with smaller values ensuring stronger privacy at the cost of noisier results. This approach is widely applied in healthcare, finance, and government, enabling data sharing for analysis and research while protecting individual records.
   d) Differential Privacy and Anonymization: A hybrid approach combines two techniques: anonymization and differential privacy. Anonymization removes obvious personal details. Differential privacy adds some controlled noise to sensitive data or models, which helps guard against risks like linking the data back to individuals or making unwanted inferences. For example, anonymized datasets can be enhanced with tools like DP-SGD (Differentially Private Stochastic Gradient Descent) during machine learning training or when creating synthetic data. This two-layer strategy not only meets privacy rules but also keeps the data useful for analysis [134].
4) **Evaluating the Trade-offs:**
   a) Balancing Privacy and Utility: Achieving an effective differential privacy implementation requires carefully balancing privacy protection

TABLE 5. Comparison of differential privacy techniques.

Technique	Description	Advantages	Disadvantages
Laplace Mechanism [128]	Adds noise drawn from a Laplace distribution, scaled by the query sensitivity and privacy budget (ϵ).	Ensures strong privacy guarantees while preserving utility for numeric queries like sums and averages.	May reduce accuracy for high-dimensional data or when frequent queries are applied.
Gaussian Mechanism [7]	Adds noise from a Gaussian distribution, suitable for approximate privacy and high-dimensional data.	Flexible and effective for high-dimensional datasets and machine learning.	Requires larger noise for stronger privacy, which can reduce utility in some cases.
Local Differential Privacy [128]	Perturbs data locally at the source, ensuring privacy without relying on a trusted aggregator.	Ideal for decentralized systems and distributed data collection.	Limited utility due to the high level of noise added to individual data points.
Exponential Mechanism [130]	Selects outputs with probabilities proportional to their utility scores, preserving categorical data privacy.	Suitable for non-numeric data and tasks like selecting the "best" option.	Computationally intensive for large datasets with many categories.
Rényi Differential Privacy [126]	Extends standard DP by using different privacy loss distributions for more flexible trade-offs.	Provides fine-grained control over privacy and utility, especially in iterative algorithms.	Complex implementation and interpretation, particularly in non-iterative settings.
Federated Differential Privacy [7]	Combines differential privacy with federated learning, ensuring privacy during collaborative model training.	Preserves privacy in distributed training environments.	Adds complexity to model training and may slightly reduce model accuracy.

with data utility. Excessive noise can compromise the usefulness of the data, limiting meaningful analysis, while insufficient noise leaves data vulnerable to privacy breaches. Evaluating these trade-offs is particularly important in contexts where user consent and understanding play a critical role, ensuring stakeholders know the implications of privacy mechanisms and their impact on data accuracy [129].

b) Ensuring Compliance: Differential privacy must also align with legal and ethical frameworks, such as GDPR and HIPAA, to ensure data-sharing practices comply with relevant regulations. This involves incorporating privacy guarantees into data workflows while adhering to stringent standards for protecting sensitive information. Compliance ensures regulatory alignment and stakeholder confidence in the privacy-preserving processes being implemented [129].

2) TOOLS & FRAMEWORKS

Differential privacy (DP) mechanisms are implemented by specialized tools and frameworks that simplify the integration of privacy-preserving techniques into different data-processing pipelines. These tools are designed to add noise, manage privacy budgets, and ensure compliance with privacy guarantees while maintaining data utility. Below, we explore some prominent tools and frameworks used in differential privacy, their features, and specific applications.

1) **Google's Differential Privacy Library [130]:**

a) Principle of Operation: Google's Differential Privacy Library is an open-source toolkit that enables DP for datasets. It provides an efficient mechanism to calculate aggregate statistics, such as counts, sums, and averages, while preserving user privacy. The library incorporates noise addition and ensures a bounded privacy budget across queries.

b) Features: Designed for Python, making it accessible for data scientists and engineers.Supports modular implementations for DP mechanisms, including Laplace and Gaussian mechanisms. It offers privacy accounting tools for tracking cumulative privacy loss over multiple queries.

c) Use Cases: Widely used in scenarios such as telemetry data analysis and privacy-aware product usage analytics. The library is part of Google's internal systems for large-scale privacy-preserving analytics.

2) **Differential Privacy Synthetic Data Toolkit [127]:**

a) Principle of Operation: This toolkit focuses on generating synthetic datasets that preserve the statistical properties of the original data while ensuring individual-level privacy. It allows organizations to share data securely without risking privacy breaches.

b) Features: Implements advanced synthetic data generation techniques using differential privacy principles. Supports interactive and batch processing modes for generating datasets. Provides configuration for adjustable privacy budgets based on specific needs.

c) Use Cases: Effective in scenarios such as training AI models on sensitive datasets or sharing anonymized data for collaboration and research.

3) **IBM Differential Privacy Toolkit [7]:**

a) Principle of Operation: IBM's toolkit provides a suite of differential privacy tools designed to integrate privacy mechanisms into AI and data analytics pipelines. It offers APIs and utilities

for applying noise to data and queries while maintaining compliance with privacy guarantees.
b) Features: Supports modular implementation of differential privacy mechanisms such as Laplace and Gaussian noise. Includes visualization tools for understanding the trade-off between privacy and utility. Offers comprehensive documentation to facilitate integration into existing machine learning workflows.
c) Use Cases: Primarily used in enterprise-level analytics and AI solutions, particularly for industries like finance and healthcare that require stringent privacy compliance.

4) **Microsoft's SmartNoise [130]:**
a) Principle of Operation: Developed in collaboration with OpenDP, SmartNoise is a library that integrates DP mechanisms into statistical analyses and machine learning workflows. It provides pre-built privacy-preserving algorithms and supports advanced DP techniques.
b) Features: Implements various DP mechanisms, including Laplace, Gaussian, and randomized response. Provides pre-configured tools for differential privacy in SQL-based data analysis. Supports integration with scalable cloud infrastructures, making it suitable for large-scale data processing.
c) Use Cases: Used extensively in academia and industry for privacy-preserving data analytics. It is especially effective for structured datasets requiring SQL-based queries.

5) **TensorFlow Privacy [7]:**
a) Principle of Operation: TensorFlow Privacy extends the popular TensorFlow library to provide differential privacy capabilities in machine learning. It introduces noise into gradient updates during training to protect sensitive data while maintaining model performance.
b) Features: Seamlessly integrates with TensorFlow, enabling easy adoption in existing projects. Supports privacy accounting tools to track cumulative privacy loss during iterative training. Includes example implementations and tutorials for quick adoption.
c) Use Cases: Ideal for training deep learning models on sensitive datasets in domains such as healthcare, finance, and personalized AI applications.

3) EVALUATING THE EFFICACY OF DIFFERENTIAL PRIVACY

1) **Privacy Protection:**
a) Re-identification Risk Assessment: Differential privacy ensures that the inclusion or exclusion of any individual in a dataset does not significantly impact the analysis results, minimizing the risk of re-identification. Techniques such as sensitivity analysis and noise calibration are used to quantify this risk and enhance protection [128].
b) Adversarial Testing: To evaluate the robustness of differential privacy mechanisms, simulated attacks, such as inference and reconstruction attacks, are used. These tests help assess how effectively the noise mechanisms protect sensitive data against adversaries with varying levels of background knowledge [126].

2) **Data Utility:**
a) Statistical Analysis: Differential privacy aims to retain the dataset's utility while introducing noise. This involves comparing key statistical properties such as distributions, means, and variances between the original and privatized datasets [130].
b) Use Case Validation: Differential privacy mechanisms are tested against specific use cases to ensure that the privatized data can still support meaningful analyses. For instance, in machine learning applications, privacy-preserving training models are evaluated for accuracy and generalizability [7].

3) **Compliance with Regulations:**
a) Legal and Regulatory Adherence: Differential privacy is increasingly recognized as a robust method for ensuring compliance with data protection laws such as GDPR and HIPAA. By providing formal privacy guarantees, differential privacy mechanisms meet legal requirements for anonymization [129].
b) Audit Trails and Documentation: Maintaining comprehensive records of the differential privacy process, including noise calibration parameters and privacy budgets, is essential for audits and compliance. Documentation helps demonstrate adherence to privacy standards while enabling reproducibility and transparency in research and applications [7].

4) **Performance Metrics:**
a) Efficiency Analysis: Evaluating the computational efficiency of differential privacy mechanisms is critical for large-scale datasets. This includes measuring processing time, noise generation efficiency, and resource utilization during implementation [126].
b) Scalability Assessment: Differential privacy mechanisms must be tested to ensure they perform effectively as data volumes increase. This involves assessing the impact of noise on larger datasets and maintaining a balance between privacy guarantees and utility. Distributed

systems employing local differential privacy are a prime example [128].

5) **User Feedback and Stakeholder Satisfaction:**
   a) Feedback from End-Users: Gathering input from data scientists, analysts, and other stakeholders ensures that differential privacy mechanisms meet the practical requirements of the end-users. For instance, insights from researchers using DP mechanisms for social network analysis reveal the importance of utility retention [130].
   b) Stakeholder Review: Collaborating with data privacy officers, legal teams, and organizational leaders is crucial to align privacy mechanisms with organizational goals and regulatory standards. Continuous review and feedback loops enhance the implementation and effectiveness of differential privacy [129].

C. PRACTICAL USES

Differential Privacy serves various practical purposes across different domains.

1) **Health Data Sharing and Research:** Healthcare institutions often require access to patient data for research, diagnostics, and public health planning. Differential privacy enables the sharing of sensitive health records without exposing individual identities. Differential Privacy's integration into federated learning models for training healthcare algorithms without violating patient privacy. For example, Genomic data is protected using DP techniques to enable correlation studies for identifying disease patterns without risking participant data exposure [7].
2) **Social Network Analysis:** Social networks generate massive amounts of sensitive data, including user interactions and preferences. By leveraging differential privacy techniques, sensitive information is protected during analyses such as subgraph counting, edge-weight estimation, and degree distribution analysis. These methods ensure that privacy is preserved without compromising the utility of the data, enabling researchers to study network behaviors in a secure manner. For example, platforms like Facebook and Twitter apply differential privacy to analyze user behavior and improve their services while safeguarding user confidentiality, thereby addressing privacy concerns in large-scale social network analysis [130].
3) **Smart Cities and IoT:** The Internet of Things (IoT) and smart cities rely on datasets collected from sensors and devices. Differential Privacy ensures these systems operate effectively while maintaining user privacy. Differential Privacy can protect sensitive location data in spatial and mapping analyses. For example, Traffic monitoring systems in smart cities analyze congestion patterns using DP to prevent revealing specific driver or vehicle information [127].
4) **Artificial Intelligence and Machine Learning:** Differential privacy is pivotal in training machine learning models on sensitive datasets. Differential Privacy mechanisms can enhance model training by protecting individual data points from exposure [126]. For example, Google's TensorFlow Privacy library employs DP mechanisms to train robust AI models while ensuring data privacy compliance [126].
5) **Census and Public Data Release:** Governments and organizations frequently publish aggregated data from censuses and surveys. Differential privacy ensures that individual responses remain confidential. Differential Privacy enables secure public data releases for policymaking and research. For example, the U.S. Census Bureau adopted DP for its 2020 census, balancing the need for accurate data with participant confidentiality [129].
6) **Financial Services and Fraud Detection:** Financial institutions rely on transactional data for fraud detection, risk assessment, and service improvements. Differential privacy is particularly effective in ensuring client data remains confidential while enabling meaningful analysis. It optimizes privacy and utility by adding carefully calibrated noise, ensuring sensitive customer information is protected during high-stakes operations such as fraud detection and credit risk modeling. For example, banks use differential privacy to detect fraudulent credit card transactions while safeguarding sensitive customer information [128].
7) **Educational Data Analysis:** Educational institutions handle vast amounts of sensitive student data, including academic records and learning behaviors. Differential privacy enables the secure analysis of this data by ensuring that individual identities remain protected, even during large-scale studies. By introducing noise to the data, DP mechanisms allow researchers to identify performance trends, analyze learning patterns, and improve educational outcomes without compromising privacy. For example, universities utilize differential privacy to anonymize student data, enabling the development of adaptive learning models and the refinement of course structures while maintaining compliance with privacy standards.mechanisms [130].

D. ADVANTAGES

Differential Privacy offers several significant advantages, particularly in privacy protection and data utility.

1) **Strong Privacy Assurance:** Differential privacy is designed to provide robust protection against privacy breaches by ensuring that the inclusion or exclusion of any individual's data does not significantly affect the output of an analysis. This property is achieved through noise injection, which makes it nearly impossible for adversaries to infer specific data points, even when armed with substantial background information [128].

2) **Enabling Secure Data Sharing and Analysis:** Differential privacy transforms data sharing by mitigating risks associated with analyzing sensitive information. Differential Privacy allows organizations to extract insights, identify trends, and develop predictive models without compromising privacy. For example, social network analysis benefits from differential privacy mechanisms by ensuring users' activities remain confidential even during large-scale data analysis. This property is essential in collaborative research, where multiple stakeholders can work on anonymized datasets to derive actionable conclusions while adhering to strict ethical and legal guidelines [130].
3) **Reducing Risks of Data Breaches:** Differential privacy significantly diminishes the risks associated with potential data breaches. By adding calibrated noise, it ensures that leaked datasets do not expose sensitive individual information. the inherent resilience of DP to adversarial inference is highlighted, providing confidence to organizations about the safety of their data, even under attack scenarios. This property is invaluable for building trust among users and stakeholders, particularly in industries like healthcare and finance, where privacy violations can have severe implications [126].
4) **Wide Applicability Across Industries:** The adaptability of differential privacy enables its use in diverse fields. From healthcare systems analyzing patient data to financial institutions processing transaction records, DP offers a customizable approach to data protection. Differential Privacy can handle a variety of data types, including structured data, unstructured text, and multimedia, making it a universal solution for privacy concerns. This flexibility supports the deployment of DP mechanisms in complex scenarios like machine learning and distributed systems, ensuring that data utility is preserved while mitigating privacy risks [7].
5) **Simplified and Cost-Effective Data Management:** The implementation of differential privacy can streamline data management practices. Differential Privacy reduces the need for stringent data-handling protocols by transforming sensitive datasets into anonymized versions. This reduces compliance costs and minimizes the complexity associated with managing personal data while maintaining analytical value [127].
Organizations leveraging DP can thus focus on extracting value from data without the overhead of developing and maintaining highly restrictive data access systems.
6) **Encouraging Innovation in Research and Technology:** Differential privacy serves as a catalyst for innovation, particularly in research and technology development. Differential privacy has been shown to empower researchers and developers to work with sensitive data while respecting privacy constraints. For instance, machine learning models trained on DP-protected data can achieve high accuracy while ensuring that individual data points remain secure [129]. This fosters advancements in areas like artificial intelligence, public health research, and policy-making, where data is critical for breakthroughs but privacy concerns often hinder progress.

E. LIMITATIONS

While differential privacy (DP) is a rigorous and widely accepted framework for privacy preservation, its practical implementation faces several challenges. Some of the limitations are discussed below:

1) **Balancing Privacy and Utility:** Achieving a balance between privacy and utility is one of the fundamental challenges in differential privacy. The privacy parameter, ϵ, governs this trade-off. A smaller ϵ ensures stronger privacy but introduces more noise into the data or query results, significantly reducing their accuracy and utility. For example, in machine learning applications, noisy data may result in models with poor predictive performance. In the context of differential privacy (DP), ϵ (epsilon) is a privacy parameter that controls the trade-off between privacy and utility. It defines the ''privacy budget'' and determines the amount of noise added to the data or query results to protect individuals' privacy. A smaller ϵ (e.g., 0.01) means stronger privacy because more noise is added to the data, making it harder to infer sensitive information. However, this comes at the cost of reduced data utility or accuracy, as the noisy data may become less useful for analysis or decision-making. A larger ϵ (e.g., 1 or higher) results in weaker privacy but better utility because less noise is added, and the data is closer to its true values. While this makes the data more useful, it also increases the risk of identifying individuals or revealing sensitive information. In essence, ϵ quantifies the trade-off between privacy protection and the quality of the data or analysis. The lower the ϵ, the more privacy is ensured, but the less accurate the data becomes [128].
2) **Parameter Sensitivity:** The choice of the privacy budget (ϵ) is a critical factor that affects the effectiveness of differential privacy mechanisms. However, determining the appropriate value for ϵ remains an open challenge with no universally accepted guidelines. Small values of ϵ guarantee strong privacy but lead to excessive noise, rendering the data less useful. Conversely, larger values compromise privacy for improved utility. The privacy budget must often be tailored to specific applications and user contexts, making it difficult to generalize. Moreover, there is no clear consensus on what constitutes an acceptable level of privacy leakage, leading to inconsistent implementations across domains.
3) **Challenges with High-Dimensional Data:** Differential privacy mechanisms struggle to handle

high-dimensional datasets effectively. High-dimensional data requires more noise to satisfy privacy guarantees, which exponentially increases the loss of utility. Furthermore, high-dimensional data often contains correlations between attributes, which can be exploited by attackers to infer sensitive information despite the application of DP. Social networks, which are inherently high-dimensional and structured, present unique challenges for DP. Current DP mechanisms often fail to balance privacy and utility when applied to such datasets, making them less effective for tasks like subgraph counting or edge weight analysis [130].

4) **Scalability and Resource Constraints:** The scalability of DP mechanisms is a significant limitation when applied to large-scale datasets or in distributed systems. Implementing DP at scale requires considerable computational resources to generate and manage noise, especially for iterative tasks such as machine learning model training or data aggregation in real-time systems. While these methods provide better privacy guarantees, the computational overhead can be prohibitive, particularly for large datasets or systems with limited processing capabilities.
5) **Susceptibility to Auxiliary Information Attacks:** Differential privacy assumes attackers can access the worst-case background knowledge. However, real-world attackers often leverage auxiliary information, such as publicly available data or externally correlated datasets, to infer sensitive details. This makes DP mechanisms vulnerable to sophisticated inference attacks, particularly in scenarios where data correlations are strong. For example, even when DP mechanisms are applied, knowledge of a user's connections or interactions may allow an attacker to re-identify them. While DP provides strong theoretical guarantees, it does not fully mitigate risks in adversarial environments.
6) **Difficulty in Comparing Privacy Mechanisms:** Comparing different privacy mechanisms operating under the same privacy budget (ϵ) is challenging. Although all mechanisms with the same ϵ provide similar worst-case privacy guarantees, their effectiveness can vary significantly. Metrics like mutual information leakage have been proposed to address this issue, as they provide a more granular understanding of privacy risks. An information-theoretic approach can help assess the performance of equivalent ϵ-privacy mechanisms. However, this approach adds complexity and computational overhead, making it difficult to adopt in resource-constrained environments [128].

We have discussed the major privacy-enhancing technologies, i.e., anonymization, encryption, synthetic data, and differential privacy elaborately. We have also compared various privacy-enhancing technologies discussed across key performance metrics like privacy guarantee, utility, scalability, security, privacy-utility trade-off, and implementation complexity, as shown in Table 6.

VII. FUTURE ASPECTS AND DEVELOPMENT

As technology continues to evolve at a rapid pace, the field of privacy engineering is poised for significant advancements in the coming years.

1) **Advanced Encryption Techniques:** With the increasing sophistication of cyber threats, encryption techniques will need to evolve to ensure robust protection of sensitive data [115]. Future advancements may include homomorphic encryption, which allows computations on encrypted data without decrypting it, thus enhancing privacy in data processing and analysis.
2) **Differential Privacy:** Differential privacy has emerged as a promising approach to protect individual privacy while allowing for meaningful data analysis [116]. Future developments may focus on refining algorithms and methodologies to achieve stronger privacy guarantees without sacrificing utility.
3) **Privacy-Preserving Machine Learning:** As machine learning and artificial intelligence applications become more prevalent, there is a growing need for privacy-preserving techniques that enable the training of models on sensitive data without compromising individual privacy [22]. Future research may lead to the development of novel techniques, such as federated learning and secure multi-party computation for privacy-preserving machine learning.
4) **Federated Learning Meets Synthetic Data:** As the demand for privacy-preserving machine learning continues to rise, the combination of Federated Learning (FL) and synthetic data holds significant potential. Federated Learning enables distributed model training while keeping data localized, minimizing privacy risks. However, challenges like malicious data contributions and model inversion attacks persist. Incorporating synthetic data generation within FL frameworks can enhance privacy further by augmenting local datasets without exposing sensitive information. Future research should focus on improving the robustness of FL through better aggregation algorithms and incentivization mechanisms, ensuring data quality, and addressing biases from synthetic data. Additionally, combining FL with differential privacy and advanced cryptographic techniques will be pivotal in securing model updates and maintaining privacy across decentralized networks [135].
5) **Privacy by Design in the Internet of Things (IoT):** The proliferation of IoT devices raises significant privacy concerns due to the vast amount of personal data they collect and process [136]. Future developments in privacy engineering will focus on integrating privacy by design principles into the development lifecycle of

TABLE 6. Comparison of different privacy-preserving techniques across key performance metrics.

Metric	Anonymization	Encryption	Synthetic Data	Differential Privacy
Privacy Guarantee	It Depends on the method of anonymization used.	Ensures data confidentiality but does not prevent metadata leakage.	Depends on the quality of the data generated and its similarity to the original data.	Provides strong privacy.
Utility	Retains utility for simple analyses but reduces utility for complex operations.	Decryption is required for utility, otherwise, it is unusable for data analytics.	Can maintain high utility if generated data is of high quality.	Can degrade significantly as privacy parameters (ϵ) increase.
Scalability	Scalable for structured data but may struggle with high-dimensional data.	Scalable, as encryption can handle large datasets efficiently.	Computationally intensive for large datasets (especially GANs).	Computationally intensive for large datasets.
Security	Susceptible to advanced attacks like linkage or background knowledge attacks.	Resistant to unauthorized access without keys.	It depends on whether synthetic data is similar enough to enable the reverse-engineering of real data.	Strong resistance to re-identification attacks if properly implemented.
Privacy-Utility Trade-off	Balances privacy and utility for small datasets may fail for large datasets.	Is Not directly applicable as encryption focuses on security rather than privacy-utility balance.	It balances privacy and utility but is prone to privacy leaks if data is too similar to the original.	Strong privacy leads to lower utility.
Implementation Complexity	Easier to implement.	Encryption libraries are well-supported but require secure key management.	It involves complex generation techniques (e.g., GANs) and evaluation frameworks.	It requires expertise to configure and implement effectively.

IoT devices, ensuring that privacy considerations are addressed from the outset.

6) **Enhancing Large Language Models (LLMs) Training with Synthetic Data:** Synthetic data helps improve existing datasets by increasing their diversity and volume. This is especially helpful when real-world data is limited, sensitive, or hard to obtain. For example, NVIDIA has developed open models to generate synthetic data to train large language models (LLMs) in industries like healthcare and finance [137].
7) **Synthetic Data for Cybersecurity:** Synthetic data can be used for cybersecurity where an organization can generate synthetic attack scenarios and network behaviors to test their defenses against cyberattacks without compromising real network data. This is especially useful for testing systems in environments where actual attack data may be scarce or too sensitive to use [138].
8) **Blockchain and Privacy:** Blockchain technology holds promise for enhancing privacy and security in various applications, including financial transactions, supply chain management, and healthcare [25]. Future advancements may involve integrating privacy-enhancing technologies such as zero-knowledge proofs and ring signatures to achieve anonymity and confidentiality in blockchain-based systems.
9) **Regulatory Compliance and Privacy Standards:** New data protection legislation, such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), will prioritize regulatory compliance and privacy requirements [20]. Future developments may involve the emergence of new privacy frameworks and standards to address evolving regulatory requirements and ensure global interoperability.
10) **User-Centric Privacy Solutions:** As individuals become more aware of their privacy rights and concerns, there will be a growing demand for user-centric privacy solutions that empower individuals to exercise control over their personal data [139]. Future developments may focus on the design and implementation of user-friendly privacy tools and mechanisms that enable individuals to manage their privacy preferences effectively.
11) **Ethical Considerations in Privacy Engineering:** As privacy engineering technologies continue to advance, it will be essential to address ethical considerations and ensure that privacy-enhancing measures are deployed responsibly and transparently [140]. Future developments may involve the integration of ethical principles such as fairness, transparency, and accountability into the design and implementation of privacy engineering solutions.

VIII. CONCLUSION

In conclusion, the field of privacy engineering offers a variety of tools and techniques to address the increasingly complex challenges of privacy in the digital age. Data anonymization, encryption, synthetic data generation, and differential privacy emerge as key pillars in the quest to protect individuals' privacy rights. Each approach brings its own set of benefits and considerations, providing a diverse toolkit for privacy practitioners and researchers. In our work, we conducted a thorough examination of data anonymization techniques, including k-anonymity, l-diversity, and t-closeness, evaluating their effectiveness against re-identification attacks and balancing data utility and privacy. For data encryption, we explored symmetric and asymmetric methods, assessed various encryption algorithms for their security and performance, and provided guidelines for selecting appropriate techniques based on data sensitivity.

In the realm of synthetic data generation, we investigated methods to create datasets that mimic real data while preserving privacy, discussed the use of synthetic data for analysis, and analyzed challenges such as model inversion attacks. In differential privacy, we have discussed various methods to generate differentially private data to be used in real-world applications to safeguard user privacy. While these technologies offer promising solutions, they are not without their limitations and ethical implications. Striking a balance between privacy preservation and data utility remains a crucial challenge. Additionally, the ever-evolving nature of technology calls for continuous innovation and adaptation in privacy engineering practices. As we move forward, collaboration between researchers, policymakers, and industry stakeholders will be essential to foster the development and adoption of robust privacy engineering solutions. Furthermore, by prioritizing privacy for all, we can create a more secure and equitable digital environment for individuals worldwide.

REFERENCES

[1] L. H. Iwaya, M. A. Babar, and A. Rashid, "Privacy engineering in the wild: Understanding the practitioners' mindset, organizational aspects, and current practices," *IEEE Trans. Softw. Eng.*, vol. 49, no. 9, pp. 4324–4348, Jun. 2023.

[2] L. F. Cranor and N. Sadeh, "A shortage of privacy engineers," *IEEE Secur. Privacy*, vol. 11, no. 2, pp. 77–79, Mar. 2013.

[3] S. Landau, "Educating engineers: Teaching privacy in a world of open doors," *IEEE Secur. Privacy*, vol. 12, no. 3, pp. 66–70, May 2014.

[4] S. Gürses and J. M. del Alamo, "Privacy engineering: Shaping an emerging field of research and practice," *IEEE Secur. Privacy*, vol. 14, no. 2, pp. 40–46, Mar. 2016.

[5] C. Prince, N. Omrani, A. Maalaoui, M. Dabic, and S. Kraus, "Are we living in surveillance societies and is privacy an illusion? An empirical study on privacy literacy and privacy concerns," *IEEE Trans. Eng. Manag.*, vol. 70, no. 10, pp. 3553–3570, Oct. 2023.

[6] S. Spiekermann, J. Korunovska, and M. Langheinrich, "Inside the organization: Why privacy and security engineering is a challenge for engineers," *Proc. IEEE*, vol. 107, no. 3, pp. 600–615, Mar. 2019.

[7] T. Zhu, D. Ye, W. Wang, W. Zhou, and P. S. Yu, "More than privacy: Applying differential privacy in key areas of artificial intelligence," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 6, pp. 2824–2843, Jun. 2022.

[8] E. Ekenstedt, L. Ong, Y. Liu, S. Johnson, P. L. Yeoh, and J. Kliewer, "When differential privacy implies syntactic privacy," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2110–2124, 2022.

[9] C. Yin, J. Xi, R. Sun, and J. Wang, "Location privacy protection based on differential privacy strategy for big data in industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 14, no. 8, pp. 3628–3636, Aug. 2018.

[10] Y. Cheng, J. Ma, Z. Liu, Y. Wu, K. Wei, and C. Dong, "A lightweight privacy preservation scheme with efficient reputation management for mobile crowdsensing in vehicular networks," *IEEE Trans. Dependable Secure Comput.*, vol. 20, no. 3, pp. 1771–1788, Mar. 2023.

[11] A. Zigomitros, F. Casino, A. Solanas, and C. Patsakis, "A survey on privacy properties for data publishing of relational data," *IEEE Access*, vol. 8, pp. 51071–51099, 2020.

[12] A. Majeed and S. O. Hwang, "When AI meets information privacy: The adversarial role of AI in data sharing scenario," *IEEE Access*, vol. 11, pp. 76177–76195, 2023.

[13] J. Powles and H. Hodson, "Google DeepMind and healthcare in an age of algorithms," *Health Technol.*, vol. 7, no. 4, pp. 351–367, 2017.

[14] S. Spiekermann and L. F. Cranor, "Engineering privacy," *IEEE Trans. Softw. Eng.*, vol. 35, no. 1, pp. 67–82, Oct. 2008.

[15] M. Boreale, F. Corradi, and C. Viscardi, "Relative privacy threats and learning from anonymized data," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1379–1393, 2020.

[16] R. D. Santos, J. Aguilar, and M. D. R-Moreno, "A synthetic data generator for smart grids based on the variational-autoencoder technique and linked data paradigm," in *Proc. Latin Amer. Comput. Conf. (CLEI)*, Oct. 2022, pp. 1–7.

[17] K. Gai, M. Qiu, and H. Zhao, "Privacy-preserving data encryption strategy for big data in mobile cloud computing," *IEEE Trans. Big Data*, vol. 7, no. 4, pp. 678–688, Oct. 2021.

[18] U. Yadav and S. Bondre, "Hybrid cryptography approach to secure the data in computing environment," in *Proc. 2nd Int. Conf. Secure Cyber Comput. Commun. (ICSCCC)*, May 2021, pp. 359–364.

[19] F. Prasser, J. Eicher, R. Bild, H. Spengler, and K. A. Kuhn, "A tool for optimizing de-identified health data for use in statistical classification," in *Proc. IEEE 30th Int. Symp. Computer-Based Med. Syst. (CBMS)*, Jun. 2017, pp. 169–174.

[20] O. Amaral, M. I. Azeem, S. Abualhaija, and L. C. Briand, "NLP-based automated compliance checking of data processing agreements against GDPR," *IEEE Trans. Softw. Eng.*, vol. 49, no. 9, pp. 4282–4303, Sep. 2023.

[21] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, "T-closeness through microaggregation: Strict privacy with enhanced utility preservation," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 11, pp. 3098–3110, Nov. 2015.

[22] J. Weng, J. Weng, J. Zhang, M. Li, Y. Zhang, and W. Luo, "DeepChain: Auditable and privacy-preserving deep learning with blockchain-based incentive," *IEEE Trans. Dependable Secure Comput.*, vol. 18, no. 5, pp. 2438–2455, Sep. 2021.

[23] C. Dimitrakakis, B. Nelson, Z. Zhang, A. Mitrokotsa, and B. I. P. Rubinstein, "Differential privacy for Bayesian inference through posterior sampling," *J. Mach. Learn. Res.*, vol. 18, no. 11, pp. 1–39, 2017.

[24] A. Altigani, S. Hasan, B. Barry, S. Naserelden, M. A. Elsadig, and H. T. Elshoush, "A polymorphic advanced encryption standard—A novel approach," *IEEE Access*, vol. 9, pp. 20191–20207, 2021.

[25] I. T. Javed, F. Alharbi, T. Margaria, N. Crespi, and K. N. Qureshi, "PETchain: A blockchain-based privacy enhancing technology," *IEEE Access*, vol. 9, pp. 41129–41143, 2021.

[26] B. Tang, C. Yang, and Y. Zhang, "A format compliant framework for HEVC selective encryption after encoding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1140–1156, Mar. 2023.

[27] A. Majeed and S. Lee, "Anonymization techniques for privacy preserving data publishing: A comprehensive survey," *IEEE Access*, vol. 9, pp. 8512–8545, 2021.

[28] S. Ji, P. Mittal, and R. Beyah, "Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 1305–1326, 2nd Quart., 2017.

[29] L. Yang, X. Chen, Y. Luo, X. Lan, and W. Wang, "IDEA: A utility-enhanced approach to incomplete data stream anonymization," *Tsinghua Sci. Technol.*, vol. 27, no. 1, pp. 127–140, Feb. 2022.

[30] X. Zhang, W. Dou, J. Pei, S. Nepal, C. Yang, C. Liu, and J. Chen, "Proximity-aware local-recoding anonymization with MapReduce for scalable big data privacy preservation in cloud," *IEEE Trans. Comput.*, vol. 64, no. 8, pp. 2293–2307, Aug. 2015.

[31] X. Zhang, L. Qi, W. Dou, Q. He, C. Leckie, R. Kotagiri, and Z. Salcic, "MRMondrian: Scalable multidimensional anonymisation for big data privacy preservation," *IEEE Trans. Big Data*, vol. 8, no. 1, pp. 125–139, Feb. 2022.

[32] M. Kanmaz, M. A. Aydın, and A. Sertbas, "A new geometric data perturbation method for data anonymization based on random number generators," *J. Web Eng.*, vol. 20, no. 6, pp. 1947–1970, 2021.

[33] X. Zhang, L. T. Yang, C. Liu, and J. Chen, "A scalable two-phase top-down specialization approach for data anonymization using MapReduce on cloud," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 2, pp. 363–373, Feb. 2014.

[34] S. Wu, X. Wang, S. Wang, Z. Zhang, and A. K. H. Tung, "K-anonymity for crowdsourcing database," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 9, pp. 2207–2221, Sep. 2014.

[35] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *Proc. 21st Int. Conf. Data Eng. (ICDE)*, 2005, pp. 217–228.

[36] H. Zhu, S. Tian, and K. Lü, "Privacy-preserving data publication with features of independent l-diversity," *Comput. J.*, vol. 58, no. 4, pp. 549–571, Apr. 2015.

[37] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discovery Data*, vol. 1, no. 1, p. 3, Jan. 2006.

[38] H. d. O. Silva, T. Basso, and R. L. de O. Moraes, "Privacy and data mining: Evaluating the impact of data anonymization on classification algorithms," in *Proc. 13th Eur. Dependable Comput. Conf. (EDCC)*, Sep. 2017, pp. 111–116.

[39] G. Loukides and A. Gkoulalas-Divanis, "Utility-aware anonymization of diagnosis codes," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 1, pp. 60–70, Jan. 2013.

[40] J. Domingo-Ferrer, J. Soria-Comas, and R. Mulero-Vellido, "Steered microaggregation as a unified primitive to anonymize data sets and data streams," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 12, pp. 3298–3311, Dec. 2019.

[41] S. P. Reiss, "Practical data-swapping: The first steps," *ACM Trans. Database Syst.*, vol. 9, no. 1, pp. 20–37, Mar. 1984.

[42] T. Neubauer and J. Heurix, "A methodology for the pseudonymization of medical data," *Int. J. Med. Informat.*, vol. 80, no. 3, pp. 190–204, Mar. 2011.

[43] B. M. L. Srivastava, M. Maouche, M. Sahidullah, E. Vincent, A. Bellet, M. Tommasi, N. Tomashenko, X. Wang, and J. Yamagishi, "Privacy and utility of X-Vector based speaker anonymization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 2383–2395, 2022.

[44] S. Zu, X. Luo, S. Liu, Y. Liu, and F. Liu, "City-level IP geolocation algorithm based on PoP network topology," *IEEE Access*, vol. 6, pp. 64867–64875, 2018.

[45] F. Praßer, F. Kohlmayer, R. Lautenschläger, and K. A. Kuhn, "ARX—A comprehensive tool for anonymizing biomedical data," in *Proc. AMIA Annu. Symp.*, Jan. 2014, pp. 984–93.

[46] M. Templ, A. Kowarik, and B. Meindl, "Statistical disclosure control for micro-data using the R packages dcMicro," *J. Stat. Softw.*, vol. 67, no. 4, pp. 1–36, 2015.

[47] D. P. Kotevski, R. I. Smee, M. Field, Y. N. Nemes, K. Broadley, and C. M. Vajdic, "Evaluation of an automated presidio anonymisation model for unstructured radiation oncology electronic medical records in an Australian setting," *Int. J. Med. Informat.*, vol. 168, Dec. 2022, Art. no. 104880.

[48] H. Mazzawi, G. Dalal, D. Rozenblatz, L. Ein-Dorx, M. Niniox, and O. Lavi, "Anomaly detection in large databases using behavioral patterning," in *Proc. IEEE 33rd Int. Conf. Data Eng. (ICDE)*, Apr. 2017, pp. 1140–1149.

[49] M. R. Fouad, K. Elbassioni, and E. Bertino, "A supermodularity-based differential privacy preserving algorithm for data anonymization," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 7, pp. 1591–1601, Jul. 2014.

[50] D. Prokhorenkov, "Toward compliance implications and security objectives: A qualitative study," in *Proc. IEEE 39th Int. Conf. Data Eng. Workshops (ICDEW)*, Apr. 2023, pp. 138–145.

[51] A. Majeed, "Attribute-centric anonymization scheme for improving user privacy and utility of publishing e-health data," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 31, no. 4, pp. 426–435, Oct. 2019.

[52] A. Gkoulalas-Divanis, G. Loukides, and J. Sun, "Toward smarter healthcare: Anonymizing medical data to support research studies," *IBM J. Res. Develop.*, vol. 58, no. 1, pp. 9:1–9:11, Jan. 2014.

[53] J. Zhou, C. Hu, J. Chi, J. Wu, M. Shen, and Q. Xuan, "Behavior-aware account de-anonymization on Ethereum interaction graph," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 3433–3448, 2022.

[54] A. Tamersoy, G. Loukides, M. E. Nergiz, Y. Saygin, and B. Malin, "Anonymization of longitudinal electronic medical records," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 3, pp. 413–423, May 2012.

[55] H. Wang, Y. Li, C. Gao, G. Wang, X. Tao, and D. Jin, "Anonymization and de-anonymization of mobility trajectories: Dissecting the gaps between theory and practice," *IEEE Trans. Mobile Comput.*, vol. 20, no. 3, pp. 796–815, Mar. 2021.

[56] M. Jegorova, C. Kaul, C. Mayor, A. Q. O'Neil, A. Weir, R. Murray-Smith, and S. A. Tsaftaris, "Survey: Leakage and privacy at inference time," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 9090–9108, Jul. 2023.

[57] J. H. Lee and S. J. You, "Balancing privacy and accuracy: Exploring the impact of data anonymization on deep learning models in computer vision," *IEEE Access*, vol. 12, pp. 8346–8358, 2024.

[58] X. Xian, T. Wu, S. Qiao, W. Wang, Y. Liu, and N. Han, "Multi-view low-rank coding-based network data de-anonymization," *IEEE Access*, vol. 8, pp. 94575–94593, 2020.

[59] J. Yoon, L. N. Drumright, and M. van der Schaar, "Anonymization through data synthesis using generative adversarial networks (ADS-GAN)," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 8, pp. 2378–2388, Aug. 2020.

[60] S. Lee and W.-Y. Shin, "Utility-embraced microaggregation for machine learning applications," *IEEE Access*, vol. 10, pp. 64535–64546, 2022.

[61] L. Xu, C. Jiang, Y. Chen, Y. Ren, and K. J. R. Liu, "Privacy or utility in data collection? A contract theoretic approach," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 7, pp. 1256–1269, Oct. 2015.

[62] D. Puthal, X. Wu, N. Surya, R. Ranjan, and J. Chen, "SEEN: A selective encryption method to ensure confidentiality for big sensing data streams," *IEEE Trans. Big Data*, vol. 5, no. 3, pp. 379–392, Sep. 2019.

[63] C. Wang, J. Lu, X. Li, P. Cao, Z. Zhou, and Q. Wen, "A personal privacy data protection scheme for encryption and revocation of high-dimensional attribute domains," *IEEE Access*, vol. 11, pp. 82989–83003, 2023.

[64] H. Li, Y. Yang, Y. Dai, S. Yu, and Y. Xiang, "Achieving secure and efficient dynamic searchable symmetric encryption over medical cloud data," *IEEE Trans. Cloud Comput.*, vol. 8, no. 2, pp. 484–494, Apr. 2020.

[65] G. Luan, A. Li, Z. Chen, and C. Huang, "Asymmetric optical image encryption with silhouette removal using interference and equal modulus decomposition," *IEEE Photon. J.*, vol. 12, no. 2, pp. 1–8, Apr. 2020.

[66] L. Harn, C.-F. Hsu, Z. Xia, and Z. He, "Lightweight aggregated data encryption for wireless sensor networks (WSNs)," *IEEE Sensors Lett.*, vol. 5, no. 4, pp. 1–4, Apr. 2021.

[67] H. Wei, C. Zhang, T. Wu, H. Huang, and K. Qiu, "Chaotic multilevel separated encryption for security enhancement of OFDM-PON," *IEEE Access*, vol. 7, pp. 124452–124460, 2019.

[68] J. Gao, H. Yu, X. Zhu, and X. Li, "Blockchain-based digital rights management scheme via multiauthority ciphertext-policy attribute-based encryption and proxy re-encryption," *IEEE Syst. J.*, vol. 15, no. 4, pp. 5233–5244, Dec. 2021.

[69] J. Li, R. Chen, J. Su, X. Huang, and X. Wang, "ME-TLS: Middlebox-enhanced TLS for Internet-of-Things devices," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1216–1229, Feb. 2020.

[70] Y. Chen, Y. Huang, J. Fu, Y. Han, K. Li, and J. Yu, "Multi wings chaotic encryption scheme for PAM-DMT-Based optical access network," *IEEE Photon. J.*, vol. 13, no. 1, pp. 1–8, Feb. 2021.

[71] S. Heron, "Advanced encryption standard (AES)," *Netw. Secur.*, vol. 2009, no. 12, pp. 8–12, Dec. 2009.

[72] H. T. Sihotang, S. Efendi, E. M. Zamzami, and H. Mawengkang, "Design and implementation of Rivest Shamir Adleman's (RSA) cryptography algorithm in text file data security," *J. Phys., Conf. Ser.*, vol. 1641, no. 1, Nov. 2020, Art. no. 012042.

[73] T. Shen, F. Wang, K. Chen, K. Wang, and B. Li, "Efficient leveled (multi) identity-based fully homomorphic encryption schemes," *IEEE Access*, vol. 7, pp. 79299–79310, 2019.

[74] C.-I. Fan, J.-C. Chen, S.-Y. Huang, J.-J. Huang, and W.-T. Chen, "Provably secure timed-release proxy conditional reencryption," *IEEE Syst. J.*, vol. 11, no. 4, pp. 2291–2302, Dec. 2017.

[75] K. Lee, "Ciphertext outdate attacks on the revocable attribute-based encryption scheme with time encodings," *IEEE Access*, vol. 7, pp. 165122–165126, 2019.

[76] K.-S. Shim, Y.-H. Kim, I. Sohn, E. Lee, K.-I. Bae, and W. Lee, "Design and validation of quantum key management system for construction of KREONET quantum cryptography communication," *J. Web Eng.*, vol. 21, no. 5, pp. 1377–1417, Jul. 2022.

[77] N. Drucker and S. Gueron, "Speeding-up P-256 ECDSA verification on × 86–64 servers," *IEEE Lett. Comput. Soc.*, vol. 2, no. 2, pp. 12–15, Jun. 2019.

[78] O. Ivanov, V. Ruzhentsev, and R. Oliynykov, "Comparison of modern network attacks on TLS protocol," in *Proc. Int. Sci.-Practical Conf. Problems Infocommun., Sci. Technol. (PIC S&T)*, Oct. 2018, pp. 565–570.

[79] Y. Shin and J. Yun, "Runtime randomized relocation of crypto libraries for mitigating cache attacks," *IEEE Access*, vol. 9, pp. 108851–108860, 2021.

[80] A. Anugurala and A. Chopra, "Securing and preventing man in middle attack in grid using open pretty good privacy (PGP)," in *Proc. 4th Int. Conf. Parallel, Distrib. Grid Comput. (PDGC)*, Dec. 2016, pp. 517–521.

[81] T. Mueller, "Let's re-sign! Analysis and equivocation-resistant distribution of OpenPGP revocations," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, Jan. 2022, pp. 34–39.

[82] *What Is.NET Framework?* Accessed: May 5, 2024. [Online]. Available: https://dotnet.microsoft.com/en-us/learn/dotnet/what-is-dotnet-framework
[83] S. M. Fawaz, N. Belal, A. ElRefaey, and M. W. Fakhr, "A comparative study of homomorphic encryption schemes using Microsoft SEAL," *J. Phys., Conf. Ser.*, vol. 2128, no. 1, Dec. 2021, Art. no. 012021.
[84] J. B. Almeida, M. Barbosa, G. Barthe, M. Campagna, E. Cohen, B. Gregoire, V. Pereira, B. Portela, P.-Y. Strub, and S. Tasiran, "A machine-checked proof of security for AWS key management service," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2019, pp. 63–78.
[85] W. El-Shafai, I. M. Almomani, and A. Alkhayer, "Optical bit-plane-based 3D-JST cryptography algorithm with cascaded 2D-FrFT encryption for efficient and secure HEVC communication," *IEEE Access*, vol. 9, pp. 35004–35026, 2021.
[86] F. Wang, B. Zhu, K. Wang, M. Zhao, L. Zhao, and J. Yu, "Physical layer encryption in DMT based on digital multi-scroll chaotic system," *IEEE Photon. Technol. Lett.*, vol. 32, no. 20, pp. 1303–1306, Oct. 15, 2020.
[87] J. Luna, A. Taha, R. Trapero, and N. Suri, "Quantitative reasoning about cloud security using service level agreements," *IEEE Trans. Cloud Comput.*, vol. 5, no. 3, pp. 457–471, Jul. 2017.
[88] *IEEE Approved Draft Standard for Interoperability of Internet Protocol Security (IPsec) Utilized Within Utility Control Systems*, Standard P2030.102.1/D1.13, 2021, pp. 1–21.
[89] Z. Gao, Q. Wu, X. Gao, Q. Li, Z. Ma, X. Wang, and Y. Qin, "25 Gb/s physical secure communication based on temporal spreading-then-random phase encryption," *IEEE Photon. Technol. Lett.*, vol. 33, no. 24, pp. 1363–1366, Dec. 15, 2021.
[90] F. Musau, G. Wang, S. Yu, and M. B. Abdullahi, "Securing recommendations in grouped P2P e-commerce trust model," *IEEE Trans. Netw. Service Manage.*, vol. 9, no. 4, pp. 407–420, Dec. 2012.
[91] W. Dai, S. Tuo, L. Yu, K. R. Choo, D. Zou, and H. Jin, "HAPPS: A hidden attribute and privilege-protection data-sharing scheme with verifiability," *IEEE Internet Things J.*, vol. 9, no. 24, pp. 25538–25550, Dec. 2022.
[92] K. Lee, "Comments on 'Secure data sharing in cloud computing using revocable-storage identity-based encryption,'" *IEEE Trans. Cloud Comput.*, vol. 8, no. 4, pp. 1299–1300, Oct. 2020.
[93] A. K. Sood, S. Zeadally, and R. J. Enbody, "An empirical study of HTTP-based financial botnets," *IEEE Trans. Dependable Secure Comput.*, vol. 13, no. 2, pp. 236–251, Mar. 2016.
[94] S. Yao, R. V. J. Dayot, I.-H. Ra, L. Xu, Z. Mei, and J. Shi, "An identity-based proxy re-encryption scheme with single-hop conditional delegation and multi-hop ciphertext evolution for secure cloud data sharing," *IEEE Trans. Inf. Forensics Security*, vol. 18, p. 3833–3848, 2023.
[95] J. Baek, E. Hableel, Y.-J. Byon, D. S. Wong, K. Jang, and H. Yeo, "How to protect ADS-B: Confidentiality framework and efficient realization based on staged identity-based encryption," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 3, pp. 690–700, Mar. 2017.
[96] C. Xu, W. Ren, L. Yu, T. Zhu, and K. R. Choo, "A hierarchical encryption and key management scheme for layered access control on H.264/SVC bitstream in the Internet of Things," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8932–8942, Sep. 2020.
[97] P. Eigenschink, T. Reutterer, S. Vamosi, R. Vamosi, C. Sun, and K. Kalcher, "Deep generative models for synthetic data: A survey," *IEEE Access*, vol. 11, pp. 47304–47320, 2023.
[98] S. Koo, C. Park, S. Lee, J. Seo, S. Eo, H. Moon, and H. Lim, "Uncovering the risks and drawbacks associated with the use of synthetic data for grammatical error correction," *IEEE Access*, vol. 11, pp. 95747–95756, 2023.
[99] H. Zhu, R. Leung, and M. Hong, "Shadow compensation for synthetic aperture radar target classification by dual parallel generative adversarial network," *IEEE Sensors Lett.*, vol. 4, no. 8, pp. 1–4, Aug. 2020.
[100] R. Houssou, M.-C. Augustin, E. Rappos, V. Bonvin, and S. Robert-Nicoud, "Generation and simulation of synthetic datasets with copulas," 2022, *arXiv:2203.17250*.
[101] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[102] A. V. Solatorio and O. Dupriez, "REaLTabFormer: Generating realistic relational and tabular data using transformers," 2023, *arXiv:2302.02041*.
[103] Z. Wan, Y. Zhang, and H. He, "Variational autoencoder based synthetic data generation for imbalanced learning," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2017, pp. 1–7.
[104] M. Rojc and I. Mlakar, "A new unit selection optimisation algorithm for corpus-based TTS systems using the RBF-based data compression technique," *IEEE Access*, vol. 7, pp. 108035–108048, 2019.
[105] D. Viana, R. Teixeira, J. Baptista, and T. Pinto, "Synthetic data generation models for time series: A literature review," in *Proc. Int. Conf. Electr., Comput. Energy Technol. (ICECET*, Jul. 2024, pp. 1–6.
[106] J. Domingo-Ferrer, K. Muralidhar, and S. Martínez, "Synthetic data generation via the permutation paradigm with optional k-anonymity," *IEEE Trans. Dependable Secure Comput.*, early access, Jan. 2, 2025, doi: 10.1109/TDSC.2024.3525149.
[107] Z. Lin, K. Ji, M. Kang, X. Leng, and H. Zou, "Deep convolutional highway unit network for SAR target classification with limited labeled training data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 7, pp. 1091–1095, Jul. 2017.
[108] M. Boedihardjo, T. Strohmer, and R. Vershynin, "Privacy of synthetic data: A statistical framework," *IEEE Trans. Inf. Theory*, vol. 69, no. 1, pp. 520–527, Jan. 2023.
[109] V. S. Chundawat, A. K. Tarun, M. Mandal, M. Lahoti, and P. Narang, "A universal metric for robust evaluation of synthetic tabular data," *IEEE Trans. Artif. Intell.*, vol. 5, no. 1, pp. 300–309, Jan. 2024.
[110] L. Fleming and O. Sorenson, "Technology as a complex adaptive system: Evidence from patent data," *Res. Policy*, vol. 30, no. 7, pp. 1019–1039, Aug. 2001.
[111] F. Zhang, C. Hu, W. Li, W. Hu, and H.-C. Li, "Accelerating time-domain SAR raw data simulation for large areas using multi-GPUs," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 9, pp. 3956–3966, Sep. 2014.
[112] A. Kothare, S. Chaube, Y. Moharir, G. Bajodia, and S. Dongre, "SynGen: Synthetic data generation," in *Proc. Int. Conf. Comput. Intell. Comput. Appl. (ICCICA)*, Nov. 2021, pp. 1–4.
[113] C. L. Abad, H. Luu, N. Roberts, K. Lee, Y. Lu, and R. H. Campbell, "Metadata traces and workload models for evaluating big storage systems," in *Proc. IEEE 5th Int. Conf. Utility Cloud Comput.*, Nov. 2012, pp. 125–132.
[114] J. Huang, Q. Huang, G. Mou, and C. Wu, "DPWGAN: High-quality load profiles synthesis with differential privacy guarantees," *IEEE Trans. Smart Grid*, vol. 14, no. 4, pp. 3283–3295, Apr. 2023.
[115] M. U. Hassan, M. H. Rehmani, and J. Chen, "Differential privacy techniques for cyber physical systems: A survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 746–789, 1st Quart., 2020.
[116] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and D. Megías, "Individual differential privacy: A utility-preserving formulation of differential privacy guarantees," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 6, pp. 1418–1429, Jun. 2017.
[117] *Synthetic Document Image With Albumentations, Imgaug and Augraphy*. Accessed: Dec. 12, 2023. [Online]. Available: https://medium.com/@ck91wei/synthetic-noisy-document-image-with-albumentations-imgaug-and-augraphy-9f1e3d99c3fc
[118] B. C. Hu, L. Marsso, K. Czarnecki, and M. Chechik, "What to check: Systematic selection of transformations for analyzing reliability of machine vision components," in *Proc. IEEE 33rd Int. Symp. Softw. Rel. Eng. (ISSRE)*, Oct. 2022, pp. 49–60.
[119] A. Gonzales, G. Guruswamy, and S. R. Smith, "Synthetic data in health care: A narrative review," *PLOS Digit. Health*, vol. 2, no. 1, Jan. 2023, Art. no. e0000082.
[120] Z. Song, Z. He, X. Li, Q. Ma, R. Ming, Z. Mao, H. Pei, L. Peng, J. Hu, D. Yao, and Y. Zhang, "Synthetic datasets for autonomous driving: A survey," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 1, pp. 1847–1864, Jan. 2024.
[121] Y. Xia, A. Arian, S. Narayanamoorthy, and J. Mabry, "RetailSynth: Synthetic data generation for retail AI systems evaluation," 2023, *arXiv:2312.14095*.
[122] S. A. Assefa, D. Dervovic, M. Mahfouz, R. E. Tillman, P. Reddy, and M. Veloso, "Generating synthetic data in finance: Opportunities, challenges and pitfalls," in *Proc. 1st ACM Int. Conf. AI Finance*, Jan. 2020, pp. 1–8.
[123] G. F. Araujo, R. Machado, and M. I. Pettersson, "Synthetic SAR data generator using Pix2pix cGAN architecture for automatic target recognition," *IEEE Access*, vol. 11, pp. 143369–143386, 2023.
[124] T. Zhang, T. Zhu, J. Li, M. Han, W. Zhou, and P. S. Yu, "Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 4, pp. 1763–1774, Apr. 2022.

[125] F. Mazzarella, M. Vespe, and C. Santamaria, "SAR ship detection and self-reporting data fusion based on traffic knowledge," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 8, pp. 1685–1689, Aug. 2015.
[126] C. Hirche, C. Rouzé, and D. S. França, "Quantum differential privacy: An information theory perspective," *IEEE Trans. Inf. Theory*, vol. 69, no. 9, pp. 5771–5787, Sep. 2023.
[127] Y. Zhang, G. Si, B. Dong, L. Chen, and X. Xu, "Privacy protection scheme for cyberspace mapping data based on differential privacy," in *Proc. IEEE 7th Inf. Technol., Netw., Electron. Autom. Control Conf. (ITNEC)*, Sep. 2024, pp. 631–634.
[128] N. Wu, C. Peng, and K. Niu, "A privacy-preserving game model for local differential privacy by using information-theoretic approach," *IEEE Access*, vol. 8, pp. 216741–216751, 2020.
[129] N. Ashena, O. Inel, B. L. Persaud, and A. Bernstein, "Casual users and rational choices within differential privacy," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2024, pp. 932–950.
[130] H. Jiang, J. Pei, D. Yu, J. Yu, B. Gong, and X. Cheng, "Applications of differential privacy in social network analysis: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 108–127, Jan. 2023.
[131] C. Dwork, "Differential privacy: A survey of results," in *Proc. Int. Conf. Theory Appl. Models Comput.* Cham, Switzerland: Springer, Apr. 2008, pp. 1–19.
[132] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, "Differentially private generative adversarial network," 2018, *arXiv:1802.06739*.
[133] M. L. Fang, D. S. Dhami, and K. Kersting, "DP-CTGAN: Differentially private medical data generation using CTGANs," in *Proc. Int. Conf. Artif. Intell. Med.* Cham, Switzerland: Springer, 2022, pp. 178–188.
[134] N. Holohan, S. Antonatos, S. Braghin, and P. M. Aonghusa, "(k,ε)-anonymity: k-anonymity with ε-differential privacy," 2017, *arXiv:1710.01615*.
[135] V. Hassija, V. Chawla, V. Chamola, and B. Sikdar, "Incentivization and aggregation schemes for federated learning applications," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 1, pp. 185–196, 2023.
[136] C. Li and B. Palanisamy, "Privacy in Internet of Things: From principles to technologies," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 488–505, Feb. 2019.
[137] *NVIDIA Releases Open Synthetic Data Generation Pipeline for Training Large Language Models*. Accessed: Dec. 23, 2025. [Online]. Available: https://blogs.nvidia.com/blog/nemotron-4-synthetic-data-generation-llm-training/
[138] D. A. Ammara, J. Ding, and K. Tutschku, "Synthetic network traffic data generation: A comparative study," 2024, *arXiv:2410.16326*.
[139] W. Jin, M. Xiao, L. Guo, L. Yang, and M. Li, "ULPT: A user-centric location privacy trading framework for mobile crowd sensing," *IEEE Trans. Mobile Comput.*, vol. 21, no. 10, pp. 3789–3806, Oct. 2022.
[140] D. Jacobs, T. McDaniel, A. Varsani, R. U. Halden, S. Forrest, and H. Lee, "Wastewater monitoring raises privacy and ethical considerations," *IEEE Trans. Technol. Soc.*, vol. 2, no. 3, pp. 116–121, Sep. 2021.

QAISER RAZI received the M.E. degree from the Birla Institute of Technology, Mesra, Ranchi, India, in 2019. He is currently pursuing the Ph.D. degree with the Department of Electrical and Electronics Engineering, Birla Institute of Technology and Science (BITS) Pilani, Pilani Campus, Pilani, India. His research interests include artificial intelligence, privacy and security, the IoT, and non-fungible tokens.

RAJA PIYUSH is a student in computer science and engineering with Kalinga Institute of Industrial Technology (KIIT), Bhubaneswar, India. His research interests include deep learning, machine learning, privacy preservation, and synthetic data generation.

ARJAB CHAKRABARTI is currently pursuing B.Tech. degree from the Kalinga Institute of Industrial Technology (KIIT), Bhubaneshwar. He is also a Research Intern with the Birla Institute of Technology and Science (BITS) Pilani, under the supervision of Dr. Vikas Hassija. He has completed a few projects in the field of machine learning and web designing. His research interests include machine learning, reinforcement learning, quantum computing, and deep learning.

ANUSHKA SINGH is currently pursuing B.Tech. degree from the Kalinga Institute of Industrial Technology (KIIT), Bhubaneswar. She is also a Research Intern with the Birla Institute of Technology and Science (BITS) Pilani, under the supervision of Dr. Vikas Hassija. Her research interests include machine learning, reinforcement learning, quantum computing, and deep learning.

VIKAS HASSIJA received the M.E. degree from the Birla Institute of Technology and Science (BITS) Pilani, Pilani, India, in 2014. He is currently an Associate Professor with the School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar, India. He was a Postdoctoral Research Fellow with the National University of Singapore (NUS), Singapore. His current research interests include blockchain, non-fungible tokens, the IoT, privacy and security, and distributed networks.

G. S. S. CHALAPATHI (Senior Member, IEEE) received the B.E. degree (Hons) in electrical and electronics engineering and the M.E. and Ph.D. degrees in embedded systems from the Birla Institute of Technology and Science (BITS), Pilani, in 2009, 2011, and 2019, respectively. He carried out his Postdoctoral Research with The University of Melbourne, Melbourne, Australia, under the supervision of Prof. Rajkumar Buyya, a Distinguished Professor with The University of Melbourne. During his Ph.D. studies, he has been a Visiting Researcher with the National University of Singapore and Johannes Kepler University, Austria. He has published in reputed journals, such as IEEE Wireless Communication Letters, IEEE Sensors Journal, and *Future Generation Computing Systems*. His research interests include UAVs, precision agriculture, and embedded systems. He is a reviewer of IEEE Internet of Things Journal and IEEE Access. He is a member of ACM.