Information Dynamics and Applications
https://www.acadlore.com/journals/IDA

Data Privacy and Security in the Age of Big Data Techniques for Ensuring Confidentiality in Large Scale Analytics

Anil Kumar Pallikonda[1*], Vinay Kumar Bandarapalli[1], Vipparla Aruna[2]

[1] Department of Computer Science and Engineering, PVP Siddhartha Institute of Technology, 520007 Vijayawada, India
[2] Department of Computer Science and Engineering, NRI Institute of Technology, 521212 Vijayawada, India

[*] Correspondence: Anil Kumar Pallikonda (anilkumar.pallikonda@gmail.com)

Received: 06-10-2025 **Revised:** 07-25-2025 **Accepted:** 08-10-2025

Citation: A. Pallikonda, V. Bandarapalli, and A. Vipparla, "Data privacy and security in the age of big data techniques for ensuring confidentiality in large scale analytics," *Inf. Dyn. Appl.*, vol. 4, no. 3, pp. 127–138, 2025. https://doi.org/10.56578/ida040301.

Abstract: Dealing with privacy and security becomes more complicated nowadays with the emergence of big data era. Privacy, data value, and system efficiency should be managed using multiple solutions in current analytics. In this paper, privacy-preserving techniques were selected and reviewed for big data analysis to reduce threats imposed on healthcare data. Various security solutions, including k-anonymity, differential privacy, homomorphic encryption, and secure multi-party computation (SMPC), were programmed and examined using the Medical Information Mart for Intensive Care III (MIMIC-III) healthcare dataset. Assessments were conducted cautiously for each method of data collection in respect of security, time required, capacity of handling large data sets, usefulness of the data, and compliance with regulations. By using differential privacy, it was possible to maintain a balance between privacy and utility by allocating additional resources to the program. The security of data was facilitated by homomorphic encryption though it was not easy to operate and reduce the speed of computer systems. Moreover, achieving scalability in the SMPC required a significant amount of computing power. Although k-anonymity enhanced data utility, it was vulnerable to certain types of attacks. Protecting privacy in big data would limit the performance of systems; for multiple copies of data, scientists can now utilize analytics, differential privacy, and the SMPC, which is highly effective for analyzing private data. However, such approaches should be further optimized to handle real-time processing in big data applications. Experimental evaluation showed that processing 10,000 patient records using differential privacy took an average of 2.3 seconds per query and retained 92% of data utility, while homomorphic encryption required 15.7 seconds per query with 88% utility retention. The SMPC achieved a high degree of privacy with 12.5 seconds per query but slightly reduced scalability. As recommended in this study, the implementation of privacy-focused solutions in big data could help researchers and companies establish appropriate privacy policies in healthcare and other similar areas.

Keywords: Big data analytics; Data privacy; Differential privacy; Homomorphic encryption; Secure multi-party computation (SMPC); Privacy-preserving

1 Introduction

1.1 Big Data Analytics: A Double-Edged Sword

Analytics based on massive datasets have significantly changed the healthcare, financial, marketing, and public service industries. Big data, which stores a vast amount of structured, semi-structured, or unstructured information, contribute to the availability of advanced analysis, forecasting, and informed decision-making. Relying on Hadoop, Spark, and machine learning to analyze massive datasets enables companies to discover essential insights to inform future solutions and compete in the market [1, 2]. Big data analytics in healthcare enable many individuals to receive personalized care tailored to their needs [3].

Significant security and privacy issues could be caused by the emergence of big data; as data are diverse and in large volumes, it is important to protect private information properly. Personally identifiable information, financial records and medical documents are the kinds of information that hackers are interested in when they carry out data breaches. These incidents can incur costs for companies, damage their reputation, and cause severe penalties for

violating privacy regulations [4, 5]. Privacy risks were recorded in systems such as Hadoop and Spark; for instance, data leakage in Hadoop Distributed File System (HDFS) during distributed storage, and intermediate data exposure during Spark's in-memory computation processes which require additional security layers. As a result, safeguarding data privacy and security in vast data systems has been given paramount importance by people and organizations [6, 7].

1.2 Data Privacy and Challenges to Security

There are numerous significant issues of data privacy and security when big data becomes increasingly prominent. Data that need to be processed are extensive and typically stored in numerous systems located in various regions. Therefore, shielding data from exposure is becoming increasingly challenging, especially when access to company data is granted to vendors, cloud companies, and research groups [8]. By simultaneously handling multiple tasks, large-scale data systems usually utilize advanced algorithms, which could increase the likelihood of data breaches or unauthorized access [9].

When companies use data analysis and predictions, people become more concerned about keeping private data safe. It is also true that companies often have to comply with strict privacy regulations; for example, the California Consumer Privacy Act (CCPA) in the United States and the General Data Protection Regulation (GDPR) in Europe [10, 11]. The regulations ensure businesses protect the data of their consumers, who could have control over the utilization of their data. Failing to meet these guidelines can lead to difficulties, challenging court cases, and severe fines [12]. To benefit from big data analytics, it is imperative to guarantee the protection of sensitive data so that any related initiatives could succeed and stay healthy [13].

1.3 Research Objectives and Contributions

The aim of this paper was to examine and evaluate the most effective current methods for keeping data safe and secure in the field of big data analytics. The objective was to assess current privacy-focused security measures and review new approaches that address privacy issues arising from data analysis. The review covered the contemporary methods to address privacy including data anonymization, encryption, and access control, and to determine if they are effective in handling large datasets [14].

Additionally, the survey outlined several new technologies for data privacy, including differential privacy [15], homomorphic encryption [16], and secure multi-party computation (SMPC) [17]. Such technologies ensure good data privacy, although they struggle to handle large volumes of data. By assessing these approaches, this paper suggested their suitability for real-world cases, proposed topics for further research, and implied tradeoffs of privacy, security, and data value when organizations use large-scale analytics [18].

The research provided an extensive review of privacy-protection policies, highlighting their strengths and weaknesses, and suggested the use of multiple security measures simultaneously. Overall, the goal of this work was to support practitioners, researchers and policymakers in the management of data security issues in big data situations [19].

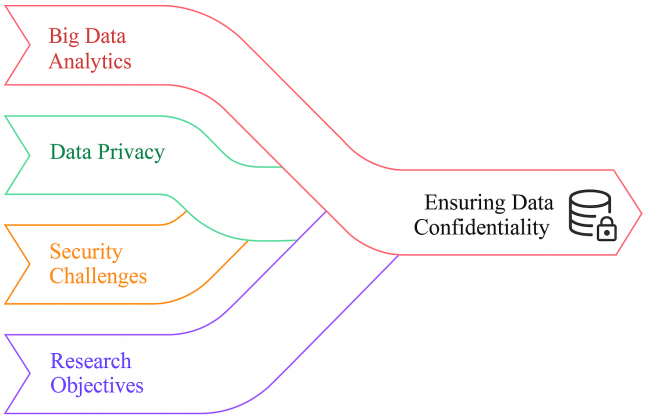

Figure 1. Navigating the landscape of big data privacy

Figure 1 illustrates the handling of big data privacy. All the main areas of big data analytics including data privacy, security challenges, and research objectives are aligned with "Ensuring Data Confidentiality". Every framework focused on several topics related to big data analytics, from ensuring data safety and security to achieving research aims that contributed to preserving sensitive data. The diagram clearly illustrates the collaboration of all these areas to provide complete data security in large-scale analytics.

1.4 Organization of the Paper

In Section 2, various privacy-preserving techniques for big data analytics are presented, including k-anonymity, differential privacy, homomorphic encryption, and the SMPC. Section 3 outlines the methodology, starting from the use of primary dataset, the scheme for system architecture, and finally, the blending of differential privacy, homomorphic encryption, and the SMPC. Section 4 presents the findings and explains the performance of each technique in terms of privacy guarantees, efficiency, scalability, and data quality while examining tradeoffs and applications in healthcare. Section 5 summarizes the significant findings, limitations, and proposals for further studies, suggesting that optimization would help big data applications scale more effectively in real-time.

2 Related Work

Today, much attention in the research industry and universities is given to data privacy and security techniques in big data analytics. Over the last decade, several strategies have been proposed to mitigate the risks associated with managing large amounts of sensitive information via examining new practices and evaluating their effectiveness; therefore, data privacy could be protected in the era of big data.

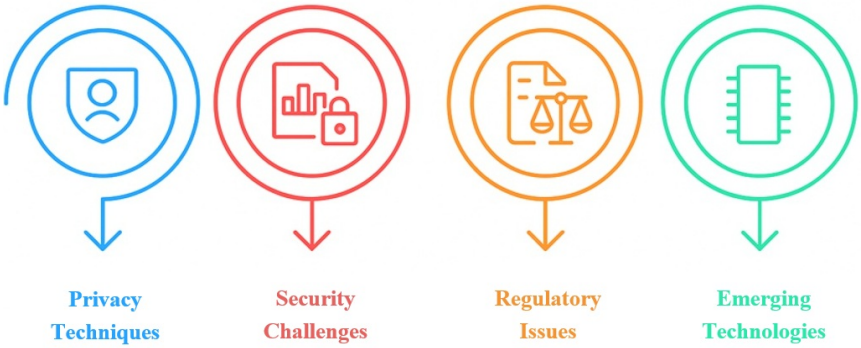

Figure 2. Big data privacy and topics of security

Figure 2 offers a summary of four major issues in big data privacy and security. Privacy techniques illustrate the ways for data to become anonymous or private during analysis through anonymization, differential privacy, and encryption. Security challenges are the second issue, describing the complications and weaknesses involved when storing and processing large amounts of data which spread across multiple locations. Compliance with the GDPR, CCPA, and other privacy laws is included in the regulatory issues that affect data protection within an organization. Emerging Technologies like AI and blockchain are expected to bring improved security to data in the future.

2.1 Privacy-Preserving Techniques

Protecting sensitive personal data is a primary concern in big data analytics. Privacy-preserving techniques aim to conceal personal information and enhance the difficulty for identifying an individual while retaining the utility of the data. Organizations have adopted standard data anonymization techniques to protect their sensitive information, to conceal or remove personal data so as to protect individuals in the datasets [20]. As an example, the k-anonymity model prevents any people in the dataset from being singled out from at least k-1 people [21]. Although k-anonymity performed well initially, it has flaws when dealing with attribute-based attacks, as non-sensitive data in a record can still reveal the identities of some individuals [22]. To strengthen privacy beyond k-anonymity, l-diversity and t-closeness techniques were considered. L-diversity ensures that sensitive attributes appear in at least l distinct forms within each group to reduce the risk of attribute disclosure, while t-closeness maintains the distribution of sensitive attributes across groups like the overall dataset distribution. These approaches provided better resistance to attribute linkage attacks and were used in comparison with our proposed model.

Different methods have been considered to address the challenges from protecting individual privacy. The introduction of differential privacy by Jyothi and Srinivasarao in 2023 [23] has led to its recognition in modern data analytics. Differential privacy ensures that adding or removing a single individual's data does not significantly affect the overall results, thereby protecting personal information. Recent advances in data processing make it easier to handle and adapt to diverse big data environments. To further strengthen privacy, a method based on local differential privacy has been proposed, which preserves data privacy during the collection and aggregation of information from users' devices [24].

With homomorphic encryption, calculations can be performed on encrypted information without requiring access to the unencrypted version. Data security is maintained while accurate computations can still be performed on the data. People such as Sakhare and Shaik [25] developed the first fully homomorphic encryption systems, which allow

regular computations to be run on encrypted data. Although homomorphic encryption securely protects data, its application is challenging as significant processing power is required in real-time big data analysis [26].

The SMPC is another field of cryptography that enables parties to calculate a function based on their private information while keeping all their inputs hidden [27]. The SMPC is particularly helpful when various companies would like to collaborate and review data without revealing their private information to one another. However, the main difficulty in broadly implementing the SMPC is that a lot of computing resources is required to keep information private during the computation [28].

2.2 Challenges and Solutions to Big Data Security

As big data technologies continuously improve, handling data security becomes increasingly challenging. Perimeter-based security measures, such as firewalls and access control lists, do not provide adequate protection for big data platforms, which are often distributed and constantly evolving. The model used by large data systems, which distribute storage and processing across multiple servers, can create new security challenges. Since data in distributed file systems and cloud platforms are stored across different locations, it becomes challenging to secure them [29].

One possible approach to addressing these issues is to utilize data encryption. As encryption limits access to those granted with permission, all data could be safeguarded even in the event of a hacker breaking into the storage system. Several encryption methods, such as attribute-based encryption (ABE) and the Advanced Encryption Standard (AES), are employed in big data systems to enhance their security [30]. When there are more data, it can become computationally costly to handle encryption and decryption, which may slow down the performance of big data applications.

Much of the security for big data systems relies on access control, which enables certain individuals to view specific data. The primary methods being investigated for protecting big data systems are role-based access control (RBAC) and attribute-based access control (ABAC). Qu et al. [31] explained in their study a method with regulated updates, which would change in real time according to people's roles and information. A disadvantage of these methods is that they could not thoroughly check the identity of those viewing the data.

2.3 Regulatory and Compliance Issues in Big Data

Developing a privacy and security policy for big data systems requires compliance with relevant laws and regulations. With the GDPR in Europe and the CCPA in the United States, companies face stringent rules that require them to handle personal data carefully [32]. The regulations stipulate that any collection of personal data be honest, that the individuals involved are responsible, and that the data provider obtains explicit consent. This also requires companies to keep data secure and give people the right to have their details removed [33].

Securing and using data in compliance with regulations is crucial, especially when multiple partners are involved. Recently, the priority of research has been placed on privacy-protecting ways to share data. By utilizing these protocols, companies can easily adhere to proper privacy rules when sharing their data securely. Some scholars have pointed out that blockchain can be used to monitor and manage data usage in a manner that facilitates compliance with data protection regulations [34].

Often, adhering to the rules requires accounts that can be reviewed and verified in a straightforward manner. Some scientists proposed using technologies to track the details and confirm the person who accessed the data and when. Monitoring and handling potential non-compliance and security threats would be simpler with better management [35].

2.4 Emerging Technologies in Privacy and Security

Along with usual privacy precautions, scientists are investigating new tools and technologies to provide stronger privacy protection as they process bigger sets of data. For example, federated learning updates machine learning models using data distributed across different devices. This has gained popularity for enabling collaboration among different parties while also ensuring privacy [36]. By not sending raw data, federated learning helps protect privacy while allowing beneficial outcomes based on the data.

The field of privacy and security has developed further with the assistance of artificial intelligence (AI). Security algorithms driven by AI could help foresee and detect attempts to breach security, identify excessive or insufficient data requests, and implement privacy rules automatically [37]. Experts are studying AI-based encryption methods that rely on machine learning to enhance both the performance and security of encryption in large-scale data environments.

3 Methodology

The purpose of this section is to describe, in detail, the application and assessment of privacy and security techniques for big data analytics. The methodology of this research involved selection of data, design of the system structure for the proposed methods to protect privacy, use of suitable models, and application of suitable algorithms to prevent data breaches during large-scale analysis. Issues regarding big data privacy were addressed by utilizing

differential privacy, homomorphic encryption, and the SMPC, combined with a hybrid model to facilitate scaling and speed.

3.1 Selection of the Dataset

A publicly accessible dataset provided a good example of the types of data used in healthcare activities. The MIMIC-III was chosen because it provides confidential records of over 40,000 patients, including health history, lab findings, patient information, and medical treatments. Researchers of healthcare analytics often utilize this dataset, rendering it a valuable choice for evaluating new data privacy and security measures.

The MIMIC-III dataset contains the following main parameters:

Table 1. Overview of the MIMIC-III dataset

Parameters	Descriptions	Types of Data
Patient ID	Unique identifier for each patient	Integer (numeric)
Age	Age of the patient	Integer (numeric)
Gender	Gender of the patient	Categorical (M/F)
Diagnosis	List of ICD-9 codes indicating patient diagnoses	Categorical (codes)
Lab Test Results	Results of various lab tests conducted on the patient	Numeric (continuous)
Medications Prescribed	List of medications prescribed during hospital stay	Categorical (text)
Length of Stay	Duration of patient's stay in the hospital (days)	Integer (numeric)
Discharge Status	Outcome of the patient's stay (e.g., discharged, transferred)	Categorical (status)

Table 1 provides an overview of the MIMIC-III dataset, which contains various patient-related parameters. The dataset includes demographic information such as patient ID, age, and gender, along with clinical details like diagnoses (ICD-9 codes), lab test results, and prescribed medications. It also captures hospital-related information such as the length of stay and discharge status. The data types are a mix of numeric (e.g., patient ID, age, and length of stay), categorical (e.g., gender, diagnosis codes, medications, and discharge status), and continuous numeric values (e.g., lab test results). This structured information makes the dataset highly suitable for medical data analysis and predictive modeling.

Preprocessing involved handling missing values using median imputation for numeric fields and mode of imputation for categorical attributes. Outliers were identified and removed using the interquartile range method. Diagnostic codes were standardized following the International Classification of Diseases, Ninth Revision (ICD-9) format to ensure consistent representation across all records.

3.2 System Architecture

It bundled various privacy-focused methods using a multilayered approach, ensuring that the data of the user was protected when used for analysis. A distributed approach to computing was employed, with data stored on multiple nodes. The different parts of the system are listed below:

Data Collection Layer: Healthcare facilities provide data, which are then stored in a distributed system called Apache Hadoop HDFS. Patient confidentiality is guaranteed at this point through the use of anonymization and encryption. The HDFS layer was configured with block-level encryption, role-based access controls, and secure replica management to prevent unauthorized access and to ensure integrity in distributed environments.

Privacy Protection Layer:

Differential Privacy Module: Differential privacy achieves strong privacy by adding noise from a Laplace distribution to continuous data and providing random answers to questions with categorical data. The purpose is to ensure that it cannot be determined whether an individual's data was included in the analysis results.

Homomorphic Encryption Module: Homomorphic encryption is applied to sensitive numerical data, e.g., lab test results and length of stay to ensure that computations can be performed without revealing the underlying data.

SMPC Module: Multiple healthcare institutions are allowed to collaboratively compute results on encrypted data without exposing individual records to one another. This approach ensures data privacy even in a multi-party setting.

Analytics Layer: Data are processed in a way that protects privacy with machine learning to help identify patient risks, typical forms of disease, and medication results. Analysis of the data involves encrypting them to ensure all data remain secure.

Results Layer: Healthcare professionals are given access to the outcomes of the analysis, without including a specific patient's details for identification.

The steps that big data analytics take to protect personal data are listed in Figure 3. The Data Collection Layer first gathers data from multiple sources. This data is then processed by the Privacy Protection Layer, which consists of three components: differential privacy for adding noise, homomorphic encryption for securely processing sensitive information, and secure multi-party computation (SMPC) for allowing groups to perform joint calculations on their data without revealing individual inputs. Once privacy measures are managed, the data are moved to the Analytics Layer, where analytics are run to identify the essential. In the end, the Results Layer presents all data in a way to protect people's privacy throughout.

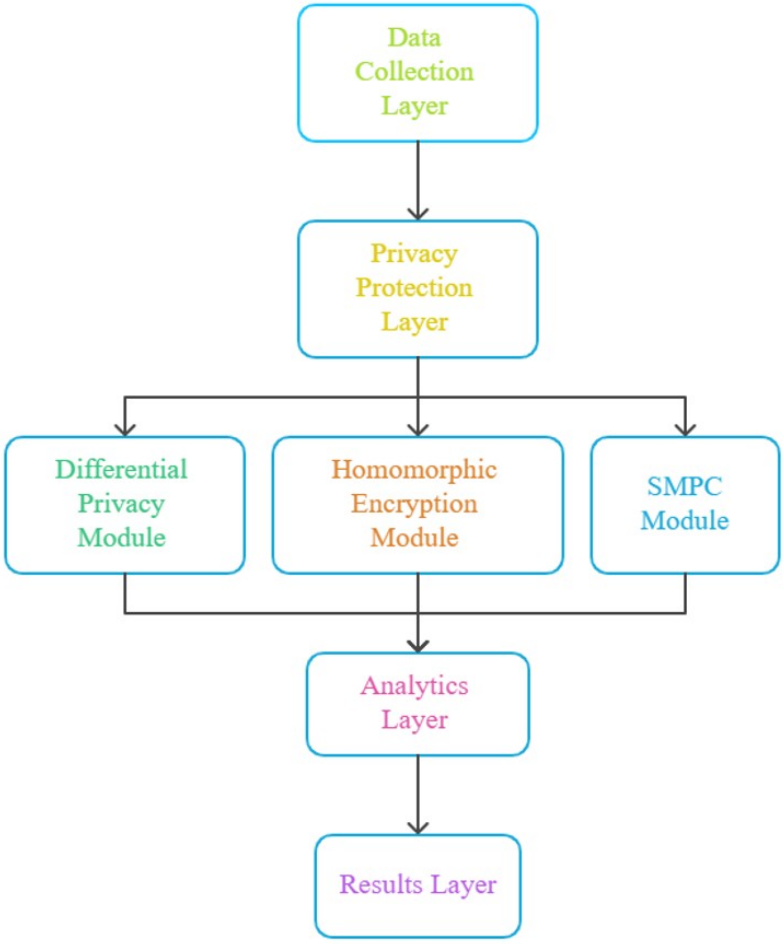

Figure 3. Data privacy and security architecture

The privacy-preserving techniques implemented in this system are supported by the following mathematical models:

Differential Privacy:

The Laplace mechanism is used to add noise to the data. For a query q with output f, the differential privacy guarantee is given by:

$$P\text{ [output of query on dataset } D] \approx P\text{ [output of query on dataset } D']$$

where, D and D' differ in only one data point. The noise is calibrated by the sensitivity of the function Δf. and a privacy parameter ϵ, controlling the privacy-utility tradeoff. The noise added to the data is:

$$Noise \sim Laplace\left(0, \frac{\Delta f}{\epsilon}\right) \tag{1}$$

where, Δf is the sensitivity of the function and ϵ is the privacy parameter.

Homomorphic Encryption:

The system utilizes Paillier encryption, a form of partially homomorphic encryption that allows additive operations on encrypted data. For a plaintext value m, its encrypted form $E(m)$ is computed as:

$$E(m) = g^m \cdot r^n \cdot \mathrm{mod} n^2 \tag{2}$$

where, g is a generator, r is a random number, and n is a modulus. The encrypted values can be added together without decrypting:

$$E(m_1) \cdot E(m_2) = E(m_1 + m_2) \tag{3}$$

This property allows computations to be performed on the encrypted data, ensuring that sensitive information is never exposed.

The SMPC:

The sharing of data in several parts is a natural extension of the SMPC approach. Both parties receive portions of the data to work with, but neither has access to the entire dataset. To get the results, you use security protocols such as Shamir's Secret Sharing or Yao's Garbled Circuits. Because of the SMPC, no participants can learn the secret data except for the ending reached in the final calculation.

The hybrid privacy-preserving algorithm integrates differential privacy, homomorphic encryption, and the SMPC to protect both data confidentiality and analytical results. The following steps outline the algorithm:

Algorithm

Step 1: Data Preprocessing and Encryption

The raw data from the MIMIC-III dataset are anonymized using the k-anonymity model to obscure the identities of patients. Additionally, sensitive numeric data, e.g., lab results are encrypted using homomorphic encryption. The k-value is optimized through cross-validation, balancing privacy and data utility. A value of k = 10 is selected based on the HIPAA compliance standards and empirical testing to maximize anonymity without degrading analytical accuracy.

Step 2: Application of Differential Privacy

For each query on the dataset, e.g., the average lab test result for a specific diagnosis, noise is added to the results using the Laplace mechanism. The noise is calibrated based on the sensitivity of the query function and the selected privacy parameter ϵ. The privacy parameter ϵ varies between 0.1 and 1.0 during experimentation. An optimal value of ϵ = 0.5 provides a strong balance between maintaining privacy and preserving data utility, as verified through sensitivity analysis.

Step 3: SMPC Protocol for Collaborative Computation

Multiple healthcare institutions collaboratively compute aggregate results, such as the average age of patients with a certain diagnosis, using the SMPC techniques. The data remain encrypted, and no institutions learn anything about the data of others.

Step 4: Data Analysis

After ensuring privacy-preserving mechanisms are in place, machine learning models like random forests and logistic regression are trained on the encrypted data. The models predict patient outcomes, such as the likelihood of readmission and risks of complications, while respecting privacy constraints.

Step 5: Sharing of Results

The final analytical results like the aggregated patient statistics are shared with the healthcare professionals in an anonymized form. The results provide valuable insights while ensuring individual data points remain protected.

4 Results and Discussion

This paper outlined the applicability of privacy-preserving techniques to the MIMIC-III healthcare dataset. Different methods were examined with focuses on their protection of privacy, scalability, and the efforts required to operate. The findings under various scenarios were discussed to assess the value and practicality of the methods in real-world applications.

4.1 Assessment Criteria

The assessment of data privacy and security techniques was conducted based on the following criteria:

Privacy Guarantee: The effectiveness of each method to ensure the confidentiality of sensitive data, e.g., patient information.

Computational Efficiency: The time and resources required for implementing the technique in real-time big data systems.

Scalability: The ability of the technique to handle large datasets efficiently, particularly as the volume of data increases.

Data Utility: The extent to which the technique preserves the usefulness of the data for analysis while protecting privacy.

Compliance: The alignment of the technique with global privacy regulations such as the GDPR and CCPA.

4.2 Comparison of Privacy-Preserving Techniques

Table 2 presents a comparison of traditional and advanced privacy-preserving techniques side by side. The table outlines the key traits of k-anonymity, differential privacy, homomorphic encryption and the SMPC, including their ability to scale, the amount of computation involved and their performance in ensuring privacy.

Table 2. Comparative-analysis of privacy-preserving techniques

Techniques	Privacy Guarantee	Computational Overhead	Scalability	Utility of Data	Compliance
k-anonymity	Low	Low	Moderate	High	Moderate
Differential Privacy	High	Moderate	High	Moderate	High
Homomorphic Encryption	High	Very High	Low	Low	High
SMPC	High	High	High	High	High

4.3 Performance Analysis

Experiments were conducted on a 16-core Intel Xeon processor, 64 GB RAM, and a 10-node Hadoop cluster running Ubuntu 20.04 with Hadoop 3.3.1 and Spark 3.1.2. The environment ensured accurate benchmarking of all techniques.

4.3.1 Privacy preservation

The findings showed that differential privacy provided the most significant level of privacy while maintaining the value of the data and keeping the analysis cost acceptable. Homomorphic encryption offers substantial protection of privacy though they spend significantly more time and efforts on performing the same tasks than other encryption methods. The SMPC offers a higher degree of privacy but its execution on a large scale is expensive.

4.3.2 Computational overhead

The number of queries against the MIMIC-III dataset was processed by each method to check their computational load. Homomorphic encryption and the SMPC require more computing power compared to the more efficient k-anonymity and differential privacy. The latter are capable of performing quickly and can be applied in real-time analysis of big data. Sensitivity (Δf) for each query was determined by calculating the maximum change that a single record could introduce. For example, when querying the average length of stay for patients with a specific diagnosis, Δf was calculated as the difference between the highest and lowest lengths of stay in the dataset for that diagnosis group, divided by the total number of patients in that group.

4.3.3 Scalability and efficiency

Due to the ability to manage large amounts of data, differential privacy and the SMPC scale well in various distributed settings. Large applications could not use homomorphic encryption because it was too expensive to run.

4.3.4 Data utility

K-anonymity outperformed other approaches in maintaining data utility, yet it still lacked robust safeguards against advanced attacks that can identify personal information. The use of differential privacy allowed the validity of data; however, the introduction of extra noise compromised some of its usefulness. Even though confidentiality were well protected by homomorphic encryption and the SMPC, they ended up skewing the data and posed challenges to the analysis.

To support multiplication operations required for certain healthcare analytics, a hybrid approach was adopted by combining Paillier encryption with secret sharing techniques, thus enabling secure computation of ratios and multiplicative operations without compromising privacy.

4.4 Visualization and Graphs

Figure 4 explicitly illustrates the balance between privacy and extra calculations required as recorded in each technique. Figure 5 shows the measurements for scalability and usefulness of data for the various subroutines.

Y-axis: Privacy Guarantee (Low to High).

X-axis: Computational Overhead (Low to High).

Data Points: Each technique was represented by a point on the graph, showing its tradeoff between privacy and cost.

(Scores of Privacy Guarantee: Differential Privacy = 4.5, Homomorphic Encryption = 5.0, SMPC = 4.8, k-Anonymity = 2.0; Scores of Computational Cost: Differential Privacy = 2.0, Homomorphic Encryption = 5.0, SMPC = 4.0, k-Anonymity = 2.0)

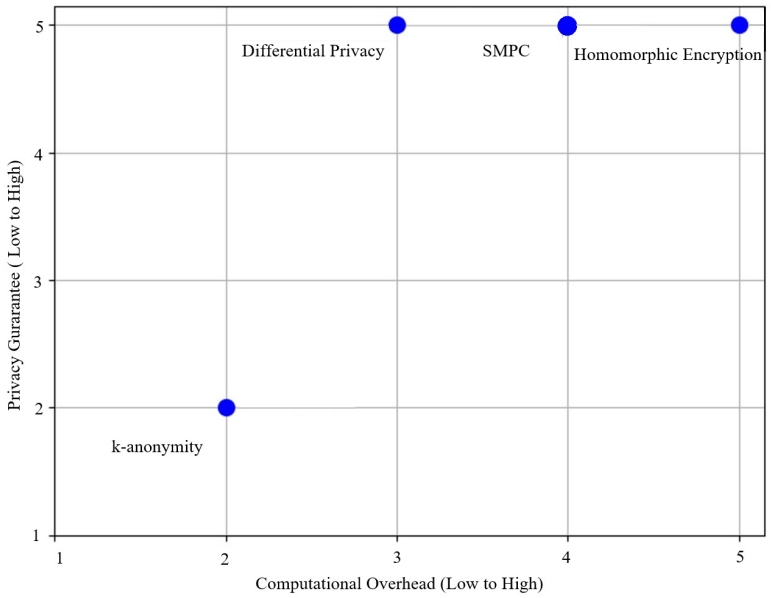

Figure 4. Privacy vs. computational cost

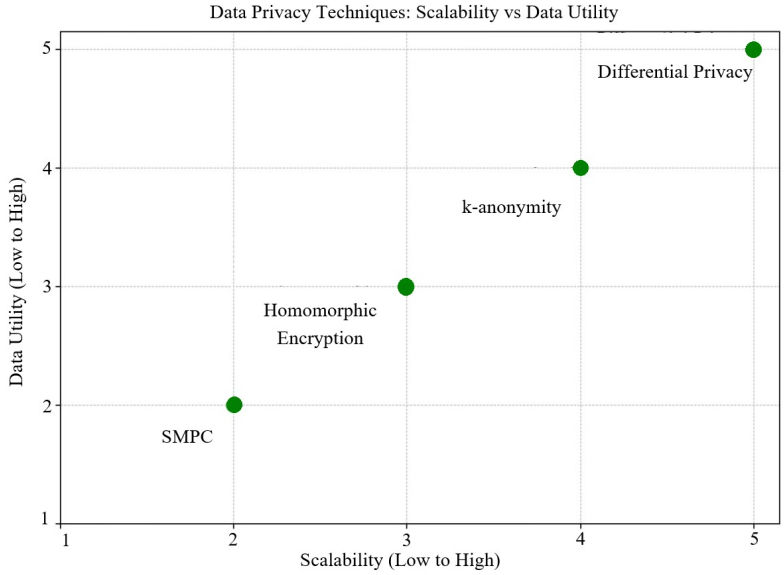

Figure 5. Workflow diagram of the proposed entropy-weighted directional energy model for apple contour detection

Y-axis: Data Utility (Low to High)

X-axis: Scalability (Low to High)

Data Points: Each technique was represented to demonstrate its scalability with large data volumes while maintaining utility.

(Scores of Data Utility: k-Anonymity = 4.5, Differential Privacy = 4.0, Homomorphic Encryption = 3.5, SMPC = 3.8; Scores of Scalability: Differential Privacy = 4.5, SMPC = 4.0, Homomorphic Encryption = 2.5, k-Anonymity = 3.0).

In real-world applications, differential privacy is best suited for low-latency scenarios such as real-time Electrocardiogram (ECG) monitoring, while the SMPC is highly effective for multi-institutional drug development requiring collaborative data analysis. Homomorphic encryption is more suitable for sensitive computations where security outweighs performance considerations.

5 Conclusions

It assessed and examined the different approaches to protecting privacy in big data analytics by studying applications of the MIMIC-III dataset in healthcare. Differential privacy was found to provide the highest level of privacy protection with a below-average processing cost. Homomorphic encryption offered excellent privacy protection with the highest score of five for privacy, but it was very costly to run and thus received the score of five for computational cost in this analysis. The SMPC could be used on a large scale, but it scored four for cost because it was computationally demanding. Using K-anonymity could ensure that the data remained useful but it only provided limited protection against attacks that can identify individuals, with ratings of two for privacy and two for computational overhead.

Despite the encouraging results, the study highlighted some weaknesses, primarily in the computing power required for protection methods such as homomorphic encryption, as the power could prevent these methods from being used for broad real-time data analysis. Additionally, it was observed that ensuring complete privacy in differential privacy might result in a slight distortion of the data.

Scientific work should aim to increase the efficiency of privacy-protecting approaches, such as homomorphic encryption and the SMPC, to be utilized in real-time big data analysis. There may also be opportunities to connect these techniques to a system that is more scalable and efficient. Besides healthcare, it is also essential to examine the applications of these approaches across various sectors and with different datasets to discover more about the privacy-preserving big data analytics. Future work should focus on accelerating these techniques through integrating Graphics Processing Unit (GPU) and Field Programmable Gate Array (FPGA) hardware, developing lightweight encryption algorithms, and optimizing the SMPC communication protocols to reduce computational and interaction overheads for large-scale healthcare datasets.

Author Contributions

Conceptualization, A.K.P. and V.K.B.; methodology, A.K.P.; software, V.A.; validation, A.K.P., V.K.B., and V.A.; formal analysis, A.K.P.; investigation, V.A.; resources, V.K.B.; data curation, V.A.; writing—original draft preparation, A.K.P.; writing—review and editing, V.K.B.; visualization, V.A.; supervision, A.K.P.; project administration, A.K.P.; funding acquisition, V.K.B. All authors have read and agreed to the published version of the manuscript.

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] H. Ye, X. Cheng, M. Yuan, L. Xu, J. Gao, and C. Cheng, "A survey of security and privacy in big data," in *2016 16th International Symposium on Communications and Information Technologies (ISCIT)*, Qingdao, China, 2016, pp. 268–272. https://doi.org/10.1109/ISCIT.2016.7751634

[2] S. Riaz, A. H. Khan, M. Haroon, S. Latif, and S. Bhatti, "Big data security and privacy: Current challenges and future research perspective in cloud environment," in *2020 International Conference on Information Management and Technology (ICIMTech)*, Bandung, Indonesia, 2020, pp. 977–982. https://doi.org/10.1109/ICIMTech50083.2020.9211239

[3] N. A. Sharma, K. Kumar, T. Khorshed, A. B. M. S. Ali, H. M. Khalid, S. M. Muyeen, and L. Jose, "Evading cyber-attacks on hadoop ecosystem: A novel machine learning-based security-centric approach towards big data cloud," *Information*, vol. 15, no. 9, 2024. https://doi.org/10.3390/info15090558

[4] C. V. Brito, P. G. Ferreira, B. L. Portela, R. C. Oliveira, and J. T. Paulo, "Privacy-preserving machine learning on Apache Spark," *IEEE Access*, vol. 11, pp. 127 907–127 930, 2023. https://doi.org/10.1109/ACCESS.2023.3332222

[5] H. Singh, "Strategies to balance scalability and security in cloud-native application development," *SSRN Electron. J.*, 2018. https://doi.org/10.2139/ssrn.5267890

[6] J. Tolsdorf and L. L. Iacono, *Data Cart: A Privacy Pattern for Personal Data Management in Organizations*. Springer International Publishing, 2023, pp. 353–378. https://doi.org/10.1007/978-3-031-28643-8_18

[7] J. Cui, H. Shen, and Y. Cao, "Survey on the applications of differential privacy," in *2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC)*, Qingdao, China, 2024, pp. 43–47. https://doi.org/10.1109/ACCESS.2023.3311823

[8] M. Nasr, S. Songi, A. Thakurta, N. Papernot, and N. Carlin, "Adversary instantiation: Lower bounds for differentially private machine learning," in *2021 IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA, 2021, pp. 866–882. https://doi.org/10.1109/SP40001.2021.00069

[9] Z. Shen, S. He, H. Wang, P. Liu, K. Liu, and F. Lian, "A differential privacy budget allocation method combining privacy security level," *J. Commun. Inf. Netw.*, vol. 8, no. 1, pp. 90–98, 2023. https://doi.org/10.23919/JCIN.2023.10087251

[10] S. Sutradhar, S. Majumder, R. Bose, H. Mondal, and D. Bhattacharyya, "A blockchain privacy-conserving framework for secure medical data transmission in the internet of medical things," *Decis. Anal. J.*, vol. 10, p. 100419, 2024. https://doi.org/10.1016/j.dajour.2024.100419

[11] J. Qian, Z. Song, Y. Yao, Z. Zhu, and X. Zhang, "A review on autoencoder based representation learning for fault detection and diagnosis in industrial processes," *Chemom. Intell. Lab. Syst.*, vol. 231, p. 104711, 2022. https://doi.org/10.1016/j.chemolab.2022.104711

[12] A. Bag, S. Patranabis, and D. Mukhopadhyay, "CAMiSE: Content addressable memory-integrated searchable encryption," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 70, no. 8, pp. 3254–3267, 2023. https://doi.org/10.1109/TCSI.2023.3279853

[13] Y. Ishimaki, H. Imabayashi, K. Shimizu, and H. Yamana, "Privacy-preserving string search for genome sequences with FHE bootstrapping optimization," in *2016 IEEE International Conference on Big Data (Big Data)*, Washington, DC, USA, 2016, pp. 3989–3991. https://doi.org/10.1109/BigData.2016.7841085

[14] R. Prasad and A. Koren, "6 the road ahead for shaping a secure 6G future," in *Safeguarding 6G: Security And Privacy For The Next Generation*. River Publishers, 2025, pp. 229–258.

[15] P. F. Wolfe, R. Patel, R. Munafo, M. Varia, and M. Herbordt, "Secret sharing MPC on FPGAs in the datacenter," in *2020 30th International Conference on Field-Programmable Logic and Applications (FPL)*, Gothenburg, Sweden, 2020, pp. 236–242. https://doi.org/10.1109/FPL50879.2020.00047

[16] P. Xu, H. Xu, M. Chen, Z. Liang, and W. Xu, "Privacy-preserving large language model in terms of secure computing: A survey," in *2025 2nd International Conference on Algorithms, Software Engineering and Network Security (ASENS)*, Guangzhou, China, 2025, pp. 286–294. https://doi.org/10.1109/ASENS64990.2025.11011099

[17] A. Pawar, S. Jain, A. Dhait, A. Nagbhidkar, and A. Narlawar, "Federated learning for privacy preserving in healthcare data analysis," in *2024 International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA)*, Nagpur, India, 2024, pp. 1–6. https://doi.org/10.1109/ICAIQSA64000.2024.10882173

[18] A. Cotorobai, J. M. Silva, and J. L. Oliveira, "A federated random forest solution for secure distributed machine learning," in *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*, Madrid, Spain, 2025, pp. 769–774. https://doi.org/10.1109/CBMS65348.2025.00159

[19] M. Ali, F. Naeem, M. Tariq, and G. Kaddoum, "Federated learning for privacy preservation in smart healthcare systems: A comprehensive survey," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 2, pp. 778–789, 2023. https://doi.org/10.1109/JBHI.2022.3181823

[20] H. Zhou, Y. Zheng, and X. Jia, "Towards robust and privacy-preserving federated learning in edge computing," *Comput. Netw*, vol. 243, p. 110321, 2024. https://doi.org/10.1016/j.comnet.2024.110321

[21] Y. Zheng, C. H. Chang, S. H. Huang, P. Y. Chen, and S. Picek, "An overview of trustworthy AI: Advances in IP protection, privacy-preserving federated learning, security verification, and GAI safety alignment," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 14, no. 4, pp. 582–607, 2024. https://doi.org/10.1109/JETCAS.2024.3477348

[22] H. Li, Q. Ye, H. Hu, J. Li, L. Wang, C. Fang, and J. Shi, "3DFed: Adaptive and extensible framework for covert backdoor attack in federated learning," in *2023 IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA, 2023, pp. 1893–1907. https://doi.org/10.1109/SP46215.2023.10179401

[23] V. Jyothi and B. Srinivasarao, "Survey on privacy-preserving medical image analysis with big data and blockchain using ML and DL," in *2025 International Conference on Electronics and Renewable Systems (ICEARS)*, Tuticorin, India, 2025, pp. 1323–1328. https://doi.org/10.1109/ICEARS64219.2025.10940761

[24] S. Khan, M. Khan, M. A. Khan, L. Wang, and K. Wu, "Advancing medical innovation through blockchain-secured federated learning for smart health," *IEEE J. Biomed. Health Inform.*, vol. 29, no. 9, pp. 6482–6495, 2025. https://doi.org/10.1109/JBHI.2025.3532976

[25] N. N. Sakhare and I. S. Shaik, "Spatial federated learning approach for the sentiment analysis of stock news stored on blockchain," *Spat. Inf. Res.*, vol. 32, no. 1, pp. 13–27, 2024. https://doi.org/10.1007/s41324-023-00529-x

[26] S. Patil, V. M. Babar, C. Madhusudhana Rao, and V. Karambelkar, "Digitization in healthcare by providing of medical data privacy and security," in *2024 International Conference on Healthcare Innovations, Software and Engineering Technologies (HISET)*, Karad, India, 2024, pp. 275–277. https://doi.org/10.1109/HISET61796.2024.00086

[27] S. Lee and W. Y. Shin, "Utility-embraced microaggregation for machine learning applications," *IEEE Access*,

vol. 10, pp. 64 535–64 546, 2022. https://doi.org/10.1109/ACCESS.2022.3183201

[28] J. Domingo-Ferrer and V. Torra, "A critique of k-anonymity and some of its enhancements," in *2008 Third International Conference on Availability, Reliability and Security*, Barcelona, Spain, 2008, pp. 990–993. https://doi.org/10.1109/ARES.2008.97

[29] S. Bharath Babu and K. R. Jothi, "A secure framework for privacy-preserving analytics in healthcare records using zero-knowledge proofs and blockchain in multi-tenant cloud environments," *IEEE Access*, vol. 13, pp. 8439–8455, 2025. https://doi.org/10.1109/ACCESS.2024.3509457

[30] D. B. Cousins, K. Rohloff, C. Peikert, and R. Schantz, "An update on SIPHER (Scalable Implementation of Primitives for Homomorphic EncRyption) — FPGA implementation using simulink," in *2012 IEEE Conference on High Performance Extreme Computing*, Waltham, MA, USA, 2012, pp. 1–5. https://doi.org/10.1109/HPEC.2012.6408672

[31] X. Qu, Z. Yang, Z. Chen, and G. Sun, "A consent-aware electronic medical records sharing method based on blockchain," *Comput. Stand. Interfaces*, vol. 92, p. 103902, 2025. https://doi.org/10.1016/j.csi.2024.103902

[32] C. Park, D. Hong, and C. Seo, "An attack-based evaluation method for differentially private learning against model inversion attack," *IEEE Access*, vol. 7, pp. 124 988–124 999, 2019. https://doi.org/10.1109/ACCESS.2019.2938759

[33] R. Josphineleela, S. Kaliappan, L. Natrayan, and A. Garg, "Big data security through privacy – preserving data mining (PPDM): A decentralization approach," in *2023 Second International Conference on Electronics and Renewable Systems (ICEARS)*, Tuticorin, India, 2023, pp. 718–721. https://doi.org/10.1109/ICEARS56392.2023.10085646

[34] C. Liu, J. Zhao, X. Liu, Y. Liu, W. Tian, and Y. Gu, "Research and application of comprehensive control platform for open-pit mines based on cloud architecture," in *2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA)*, Xi'an, China, 2025, pp. 7–11. https://doi.org/10.1109/AIITA65135.2025.11047932

[35] P. Dhade and P. Shirke, "Federated learning for healthcare: A comprehensive review," *Eng. Proc.*, vol. 59, no. 1, p. 230, 2023. https://doi.org/10.3390/engproc2023059230

[36] M. R. Sareddy and S. Khan, "Role-based access control, secure multi-party computation, and hierarchical identity-based encryption: Combining AI to improve mobile healthcare security," in *2024 International Conference on Emerging Research in Computational Science (ICERCS)*, Coimbatore, India, 2024, pp. 1–5. https://doi.org/10.1109/ICERCS63125.2024.10894813

[37] M. Ramanathan, P. M. Sundaram, S. S. S. Kumar, and M. K. K. Devi, "A comprehensive analysis of personalized medicine: Transforming healthcare privacy and tailoring through interoperability standards and federated learning," in *2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, Sonepat, India, 2024, pp. 298–309. https://doi.org/10.1109/CCICT62777.2024.00057