Proceedings of the 9th International Conference on Electronics, Communication and Aerospace Technology (ICECA-2025)
DVD Part Number: CFP25J88-DVD; ISBN: 979-8-3315-9928-7

Securing Sensitive Health Data: A Survey of Privacy-Preserving Approaches in Data Science

Anagha Achuthan
Department of Computer Science And Engineering
Vels Institute of Science, Technology And Advanced Studies
anagha.achuthan@gmail.com

Dr. A.Saritha
Department of Computer Science And Engineering
Vels Institute of Science, Technology And Advanced Studies
saritha.se@vistas.ac.in

***Abstract*— Privacy-preserving data science in healthcare has become increasingly important as the industry shifts towards data-driven decision-making. Privacy refers to the protection of sensitive patient information from unauthorized access, misuse, or disclosure. Maintaining confidentiality while allowing meaningful analysis poses a significant challenge as the healthcare data volume and complexity keep rising. Traditional privacy protection methods such as encryption, anonymization, and access control have been widely used but often fail to maintain data utility or prevent advanced re-identification attacks. Emerging privacy-preserving techniques such as federated learning, homomorphic encryption, differential privacy, and secure multi-party computation provide innovative solutions by enabling collaborative model training and analysis without compromising individual data privacy. These models are important in the healthcare domain, where decentralized data and strict privacy regulations limit data sharing. The objective of this study is to systematically review and analyze these privacy-preserving techniques in healthcare, identify their strengths and limitations, and suggest directions for future improvement. The findings indicate that while these methods enhance confidentiality and enable distributed analytics, they still face challenges related to computational overhead, interoperability, and scalability. This literature-based study explores the development of privacy-preserving data science methods, discusses the limitations of existing techniques, and highlights the need for robust, secure, and ethical data science frameworks in healthcare.**

Keywords—Data Mining, Healthcare Sector, Data Analytics, Federated Learning, Homomorphic encryption, Data Security, Encryption

I. INTRODUCTION

In the digital transformation era, the healthcare sector is experiencing an incredible increase in data generation and acquisition. This extensive and intricate data, also known as Big Data, offers significant opportunities for the improvement of patient care, disease prediction, lowering expenses and enabling early treatments. The diversity and rapid rate of healthcare data provide considerable challenges for traditional data processing systems. Big data analytics is revolutionizing the healthcare sector, empowering medical professionals with insights that were previously inaccessible. However, the management of such sensitive information also raises significant issues about privacy, security, and ethical utilisation.

As healthcare systems use advanced analytical techniques, data mining becomes an effective tool for extracting valuable insights from large datasets. It is applied across different fields, such as public health surveillance, therapy optimization, and medical diagnostics. However, as the data mining approaches grow more sophisticated, they also pose higher risks to privacy. Leveraging these analytical tools without compromising the confidentiality of personal health information is the difficult part. Techniques such as classification, clustering, and statistical modelling are adopted to ensure that the sensitive data attributes are protected from the unauthorized disclosures throughout the analytical process.

The emergence of digital records and integrated healthcare systems has increased the necessity for strong privacy-preservation measures [1]. Traditional privacy preservation protection methods such as anonymization, often reduce the data utility. Although data mining and machine learning (ML) have emerged as crucial techniques in extracting knowledge from the healthcare data, they also introduce the risks of privacy breaches. To address these gaps, innovative privacy preserving approaches have emerged, balancing the data utility and the privacy safeguards. Some of them are Federated Learning (FL), Differential Privacy (DP), Secure Multi-Party computation (SMPC) and Homomorphic Encryption (HE). Fig. 1 illustrates the privacy preservation of Internet of Things (IoT) healthcare data.

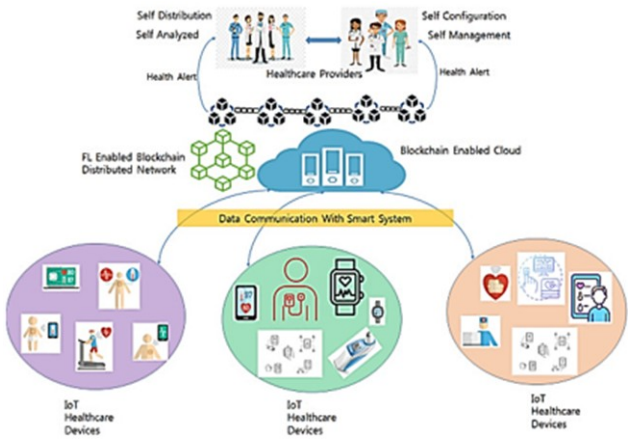

Fig. 1. Privacy preservation of IoT healthcare data [2]

FL allows decentralised model training across healthcare organisations without sharing the raw data. Each organisations use its own private data to train a local model, and only encrypted model updates are sent to a central server for aggregation. HE allows computations to be performed on encrypted data without decrypting it first. This makes it possible to analyse data in healthcare context while preserving confidentiality, making it suitable for sensitive medical records. A cryptography method called SMPC enables to allows multiple parties to jointly compute a function without disclosing specific datasets to third parties. DP introduces a controlled amount of statistical noise to query results or data outputs, making it difficult to identify any individual data record. As healthcare systems move towards integrated and

intelligent infrastructures, adopting these methods is essential to protect patients' rights while unlocking the full potential of data analytics [3]. Fig. 2. provides a concise graphical overview of the survey, mapping the progression from the problem context and research objectives to the core techniques reviewed (Federated Learning, Homomorphic Encryption, Differential Privacy, and Secure Multi-Party Computation). It illustrates the systematic review method used to collect, screen, classify, and compare studies, clarifying how the analysis is organized. The flow culminates in the synthesis of findings and future directions, helping readers quickly grasp the scope and structure of the paper.

Fig. 2. Graphical overview of the survey

Ensuring data security and privacy in healthcare requires a multilayered framework combining encryption, decentralized learning, and secure key management. Security is achieved through HE and SMPC, which allow computations on encrypted data without revealing patient information, and FL, which decentralizes model training so raw data never leave local servers. Privacy is further enhanced through DP, which injects statistical noise into results to prevent individual re-identification, while hybrid FL-DP or FL-HE schemes balance privacy and analytical accuracy. The encryption process is improved using lightweight Fully Homomorphic Encryption (FHE) and Elliptic-Curve Cryptography (ECC) to reduce computational cost and latency. Secure keys are generated via Elliptic-Curve Diffie-Hellman (ECDH) protocols using cryptographically secure randomization and managed hierarchically to avoid single-point failures. Collectively, these mechanisms ensure confidentiality, integrity, and scalability in privacy-preserving healthcare data science. The main objectives of this study are:

- To systematically review and analyse existing privacy-preserving techniques in healthcare, focusing on methods such as FL, DP, SMPC and HE.
- To identify current trends, challenges and gaps in privacy-preserving approaches, providing insights for future developments in secure and ethical health care data science.

II. Literature Review

This study adopts a systematic literature review methodology to analyze privacy-preserving techniques in healthcare data science. The process involved identifying relevant peer-reviewed journal papers published between 2020 and 2025 from digital databases such as IEEE Xplore, SpringerLink, Scopus, and ScienceDirect. Studies focusing on FL, HE, DP, and SMPC were shortlisted based on their contribution to data security and privacy enhancement. The selected papers were then classified according to methodology, advantages, limitations, and performance metrics. A comparative analysis was conducted to derive insights, challenges, and emerging research directions. The overall process is illustrated in Fig. 3, which outlines the sequential stages of study selection, analysis, and conclusion formulation.

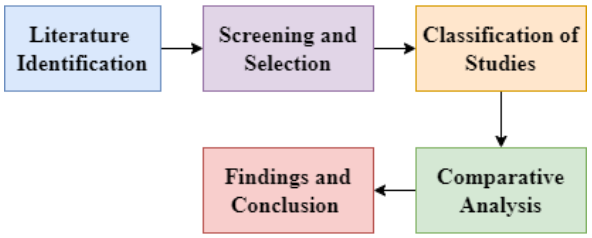

Fig. 3. Research methodology framework

A. Homomorphic Encryption for Privacy Preservation in Healthcare

HE is an advanced cryptographic method that enables computations to be carried out directly on encrypted data without needing to decrypt it, as shown in Fig. 4. This unique capability ensures that sensitive data remains confidential during processing, making HE especially valuable in fields such as finance, healthcare, and confidential research. It allows mathematical operations to be performed on encrypted information, with the final result also encrypted only the output is decrypted when needed. The main goal of HE is to protect data privacy while allowing meaningful analysis or computation, ensuring both security and integrity throughout the process.

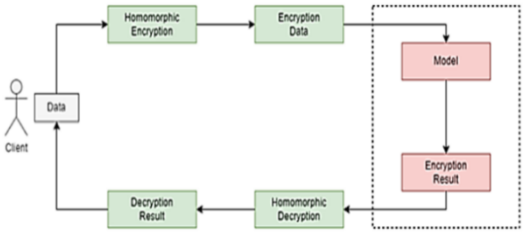

Fig. 4. Homomorphic Encryption [4]

Naresh et al. (2025) [5] proposed a HE-Driven Logistic Regression (HELR) framework for privacy-preserving heart disease prediction using the Cheon-Kim-Kim-Song (CKKS) encryption scheme. The model was implemented using the TenSeal and Torch libraries and evaluated on encrypted healthcare datasets sourced from Kaggle, using varying polynomial degrees. The model was tested across three datasets and compared with Support Vector Machines (SVM). Results showed that the model achieved accuracy within 1–3% of its non-encrypted counterpart while outperforming SVM. The system also demonstrated strong resilience against various privacy attacks, including poisoning, evasion, member inference, model inversion, and model extraction. Despite its effectiveness, the model reported limitations such as computational overhead and challenges in scaling the model to larger datasets.

Babu et al. (2025) [6] addressed the critical importance of safeguarding healthcare data privacy in distributed ML systems. The model balanced confidentiality, security, and data insight extraction by employing FHE as a transformative solution. The methodology integrated regression techniques, collaboration mechanisms, access controls, and benchmarking to secure sensitive health data and prevent breaches while enabling collaborative analysis. Experimental evaluations

demonstrated that the model delivered improved accuracy and precision compared to existing approaches, all while maintaining comparable time complexity. Although the model optimized performance and enhanced data security, it exhibited challenges in managing dynamic, decentralized healthcare data environments, suggesting potential limitations in scalability and practical deployment across diverse settings.

Panzade et al. (2024) [7] introduced MedBlindTuner, a privacy-preserving fine-tuning model that combined FHE with a Data-Efficient image Transformer (DEiT) to enable secure ML on encrypted medical images. The model addressed privacy concerns related to outsourcing ML tasks, particularly in scenarios involving sensitive data like chest X-rays. MedBlindTuner allowed model training directly on FHE-encrypted images, eliminating the need to decrypt data during computation. The methodology ensured end-to-end privacy while maintaining high model performance. Experimental evaluations demonstrated that the model achieved accuracy comparable to models trained on unencrypted data, with only minimal deviations. The model was inefficient in handling complex medical image datasets, highlighting limitations in its scalability and generalizability.

Sarkar et al. (2023) [8] addressed the challenge of privacy-preserving cancer type prediction using HE on a dataset comprising over 2 million genetic mutations from 2,713 patients. The model encoded somatic mutations based on their biological impact and applied statistical tests for feature selection. A logistic regression model was then optimized using a fast matrix multiplication algorithm designed for HE to enable high throughput and low latency. The model achieved a micro-average Area Under the Curve (AUC) of 0.98 and improved accuracy from 70.08% to 83.61%, with encrypted inference performed in approximately 1 second per patient. In comparison with plaintext models built on The Cancer Genome Atlas (TCGA) database, this HE-based approach demonstrated superior performance. However, limitations included high computational overhead with high-dimensional genomic data and inefficiencies of standard matrix multiplication algorithms for smaller matrices under HE due to large ciphertexts and poor cache utilization. Communication-efficient protocols were avoided due to their high latency.

Ali et al. (2023) [9] proposed a blockchain-integrated privacy-preserving framework for IoT-based healthcare applications using HE and smart contracts. The model employed a fully HE schemes based on elliptic curve cryptography (FHE-ECC) for secure data collection and storage, selecting optimal group keys during ECC-based encryption to reduce computational time and memory usage. Medical data were encrypted before being stored on the blockchain and decrypted only when needed for prediction tasks. The model incorporated a deep learning-based Optimized Deep Neural network–Gated Recurrent Unit (ODNN-GRU) model, to enable accurate medical predictions. Performance optimization was achieved by fine-tuning model parameters. Experimental results showed that the model achieved superior accuracy, outperforming traditional methods such as Merkle Tree, DNN, GRU, and DNN-GRU by 7.52%, 5.37%, 3.11%, and 1.17% respectively. However, the model identified limitations including the lack of real-world deployment to evaluate practical feasibility, usability improvements for end-user adoption, and extensive security assessments to address potential vulnerabilities. The summary of the recent works in the privacy preservation in healthcare using homomorphic encryption is given in Table I.

TABLE I. SUMMARY OF THE RECENT WORKS IN PRIVACY PRESERVATION IN HEALTHCARE USING HE

Author (s)	Methodology	Advantages	Disadvantages
Naresh et al. (2025) [5]	HELR	• Achieved accuracy within 1–3% of its non-encrypted counterpart while outperforming SVM. • Strong resilience against various privacy attacks.	• High computational complexity. • Challenges in scaling the model to larger datasets.
Babu et al. (2025) [6]	FHE	• Model optimized performance and enhanced data security. • Model combined robust encryption with effective collaboration measures.	• Challenges in managing dynamic, decentralized healthcare data environments. • Limitations in scalability and practical deployment across diverse settings.
Panzade et al. (2024) [7]	MedBlindTuner	• Achieved accuracy comparable to models trained on unencrypted data, with only minimal deviations.	• Inefficient in handling complex medical image datasets. • Limitations in its scalability and generalizability.
Sarkar et al. (2023) [8]	FHE	• Achieved a micro-average AUC of 0.98 and improved accuracy from 70.08% to 83.61%, with encrypted inference performed in approximately 1 second per patient. • Superior performance.	• Limitations included high computational overhead with high-dimensional genomic data and inefficiencies of standard matrix multiplication algorithms for smaller matrices under HE. • Communication-efficient protocols were avoided due to their high latency.
Ali et al. (2023) [9]	FHE-ECC	• Achieved superior accuracy, outperforming traditional methods such as Merkle Tree, DNN, GRU, and DNN-GRU by 7.52%, 5.37%, 3.11%, and 1.17% respectively.	• Lack of real-world deployment to evaluate practical feasibility. • Requirement for usability improvements to enhance for end-user adoption. • Lack of extensive security assessments to address potential vulnerabilities.

B. Federated Learning for Privacy Preservation in Healthcare

FL is a transformative approach in healthcare that enables collaborative ML across multiple institutions without sharing raw patient data, as shown in Fig. 5. By allowing models to train locally on devices such as IoT sensors, wearables, and hospital databases, FL preserves data privacy while leveraging diverse datasets. This decentralized method is especially vital

in healthcare, where legal and ethical constraints often limit centralized data access. FL supports applications like remote monitoring, personalized medicine, and clinical decision-making, all while ensuring sensitive data remains secure. It also helps build trust in AI-powered healthcare solutions by safeguarding patient confidentiality and mitigating risks from malicious data manipulation.

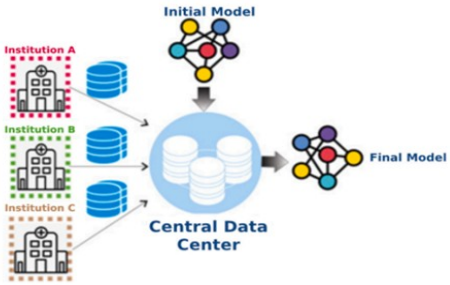

Fig. 5. Federated Learning in Healthcare [10]

Meduri et al. (2025) [11] introduced a FL framework for privacy-preserving analysis of Electronic Health Records (EHRs), aimed at improving rare disease research. The model enabled multiple institutions to collaboratively train ML models without directly sharing sensitive patient data, maintaining compliance with Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR). Various algorithms, including Logistic Regression, Decision Trees, Support Vector Classifiers, Random Forests (RF), and Stacking Classifiers, were evaluated. The RF classifier showed the best performance with 90% accuracy and an F1 score of 80%, demonstrating robustness on complex, imbalanced datasets. The model incorporated DP and SMPC to strengthen privacy and efficiency. A key limitation of the model was its focus on specific use cases, leaving the framework's effectiveness in broader healthcare areas like chronic disease management and personalized care largely unexplored. Additionally, the current privacy-preserving mechanisms require enhancement to handle more diverse data types, such as genomic and wearable health data, while maintaining robust privacy protection.

Yang et al. (2024) [12] proposed FL for privacy-preserving collaboration in drug development. The model detailed core FL mechanisms such as model parameter updating, momentum restart strategies, DP and HE. The model enabled decentralized model training across multiple institutions, preserving data locality and confidentiality. Experimental results demonstrated that momentum restart significantly improved convergence and model stability, while selective sharing of model parameters optimized the balance between privacy and performance. Real-world examples, including NVIDIA Clara and COVID-19 resource prediction, showcased the efficiency of FL in multi-party collaboration. The model highlighted FL's capacity to enhance model accuracy and accelerate drug development without compromising patient privacy. However, a notable limitation was the lack of detailed information about the dataset used and its diversity, which affects the generalizability and model robustness in broader clinical or pharmaceutical applications.

A FL model for privacy preserving medical image analysis was developed by Muthalakshmi et al. (2024) [13]. The model utilised DL approaches that were trained collaboratively by multiple organisations using their local datasets without sharing the raw data. In order to ensure the data security, the model employed DL, advanced encryption techniques and secure communication protocols. The system was evaluated using the real-world medical imaging datasets from chest, breast and abdomen scans. The model achieved a high overall accuracy of 98.6%. Federated averaging and HE enhanced the system's privacy capabilities. However, the model faced limitations such as the convergence speed was normal rather than optimized, the average communication cost was high and training time reached 16 hours.

Markkandan et al. (2024) [14] introduced a privacy preserving Collaborative Medical Diagnosis (CMD) architecture that integrates FL with a Partially Homomorphic Cryptosystem (PHC), which secured data during transmission and a Residual learning-based Deep Belief Network, which improve the classification accuracy. The method provided decentralised model training among healthcare organisations without the need to exchange raw data and thus preserving patient privacy. Experimental findings showed a 30% of decrease in computational overhead and 10% increase in diagnostic accuracy. Classification and encryption times were 1000ms and 150ms respectively. The model balanced the security as well as performance effectively. The limitation of the model was the lack of evaluation across the heterogenous healthcare environments and its low computational speed.

Mohammed Abaoud et al. (2023) [15] proposed a framework for privacy-preserving FL models tailored for healthcare applications. The approach enabled healthcare organizations to collaboratively train ML models on decentralized data while preserving patient confidentiality. Advanced privacy techniques, including SMPC and DP, were integrated during model aggregation to protect sensitive information. Comprehensive simulations and numerical analyses demonstrated that the FL models achieved an average accuracy of 97.69%, significantly outperformed existing approaches in offering more utility and assuring efficient privacy protections. Enhanced computational efficiency led to faster data analysis and reduced processing workload. The method demonstrated the effectiveness of secure and privacy-preserving collaboration on healthcare data, demonstrating its practicality and efficacy. Limitations included challenges in deploying the approach across heterogeneous healthcare environments.

Zhang et al. (2022) [16] developed a FL framework for privacy preservation in IoT-based healthcare applications. A weighted average technique predicated on data quality was introduced to replace the conventional weight calculation method reliant on data volume. A masking approach utilizing HE and SMPC was presented for FL. The model also examined classification accuracy and privacy protection measures. The simulation results suggested that the model was capable of detecting lesion cell types with an accuracy of 76.9%. Despite the results, limitation included the lack of support for the heterogeneous client environments with constrained hardware and the exclusion of asynchronous FL scenarios. Additionally, the model did not address the security threats from potentially malicious servers or tampered aggregated models, which affected the system robustness.

Li et al. (2021) [17] proposed ADDETECTOR, a privacy-preserving smart healthcare system designed for low-cost Alzheimer's disease (AD) detection using audio data collected from IoT devices. The system employed a three-layer

architecture (user, client, cloud) to safeguard privacy at data, feature, and model levels. FL was integrated to retain user data integrity, while differential privacy DP was applied to protect sensitive features. An asynchronous privacy-preserving aggregation module further secured model updates in the FL process. The model was evaluated using 1,010 AD detection trials from 99 participants, including both AD and healthy users. Results demonstrated a high detection accuracy of 81.9% with a low time overhead of 0.7 seconds, even with all privacy mechanisms in place. One limitation was the limited dataset size, which impacted the system's ability to generalize effectively to larger populations or more complex clinical scenarios. The summary of the recent works in the privacy preservation in healthcare using federated learning is given in Table II.

TABLE II. SUMMARY OF THE RECENT WORKS IN PRIVACY PRESERVATION IN HEALTHCARE USING FL

Author (s)	Methodology	Advantages	Disadvantages
Meduri et al. (2025) [110]	FL	• RF classifier showed the best performance with 90% accuracy and 80% F1 score.	• Focused only on specific use cases, not generalized. • Limited evaluation in areas like chronic disease management and personalized care. • Inefficient in handling diverse data types such as genomic and wearable health data, while maintaining robust privacy protection.
Yang et al. (2024) [12]	FL	• Decentralised model training. • Improved convergence and model stability.	• Lack of detailed information about the dataset used and its diversity.
Muthalakshmi et al. (2024) [13]	FL	• Achieved high accuracy of 98.6%.	• Limitations such as the convergence speed was normal rather than optimized. • Average communication cost was high.
Markkandan et al. (2024) [14]	FL-PHC	• Reduced computational overhead. • Decentralised training.	• Lack of evaluation across the heterogenous healthcare environment. • Low computational speed.
Mohammed Abaoud et al. (2023) [15]	FL	• Achieved high accuracy of 97.69%. • Faster data analysis.	• Limitations included challenges in deploying the approach across heterogeneous healthcare environments.
Zhang et al. (2022) [16]	FL	• Data quality-based weighting. • Effective lesion detection.	• Lack of support for the heterogeneous client environments with constrained hardware and the exclusion of asynchronous FL scenarios. • Model did not address the security threats from potentially malicious servers.
Li et al. (2021) [17]	FL	• Low-cost detection. • High accuracy and low time overhead.	• Limited dataset.

C. Differential Privacy for Privacy Preservation in Healthcare

DP is a powerful method for protecting individual data during sharing and analysis. It achieves this by adding carefully calibrated noise to the data or query results, ensuring that the inclusion or removal of a single record does not significantly affect the analysis. This prevents the identification of any individual within the dataset. DP is widely applied in statistics and ML to maintain privacy while preserving data utility. However, it requires a delicate balance, excessive noise diminish data accuracy, while insufficient noise fail to protect individual privacy.

Alsenani et al. (2025) [18] developed a FL framework integrating DP for secure patient activity monitoring using wearable device data. The methodology applied privacy-preserving mechanisms including Laplace, Gaussian, and Exponential noise on statistical functions. The Human Activity Recognition dataset was used to evaluate performance. The model demonstrated near-equal clustering utility to non-DP setups, outperforming FL in clustering quality. Feature-specific scaling-maintained privacy-utility trade-offs, effectively protecting sensitive data. Results showed efficient scalability and stable performance with increasing clients. The model also surpassed other privacy-preserving clustering techniques in accuracy and stability. However, limitations included assumptions of honest participants, leaving the system vulnerable to adversarial clients submitting manipulated summaries. The lack of adversarial resilience and hybrid privacy mechanisms marked areas for improvement. Despite these, the model proved effective for privacy-preserving, scalable healthcare analytics.

Wang and Li (2023) [19] proposed a privacy-preserving medical data collection method, Medical Data Local Differential Privacy (MDLDP), combining local differential privacy with Count Sketch for use in the Internet of Medical Things (IoMT). The methodology involved random sampling where each user perturbed a single symptom before data upload. Count Sketch was used for aggregation, enabling estimation of symptom frequency and mean occurrence. The algorithm was evaluated using a real medical dataset and compared with existing local differential privacy techniques. Experimental results demonstrated that MDLDP ensured unbiased estimation and preserved data correlations while maintaining a high level of usability and accuracy for key-value medical data. The algorithm outperformed other approaches in preserving privacy without significantly compromising data utility. However, limitations included the lack of improved encoding methods and optimization of perturbation techniques to further enhance performance. These challenges affected the scalability and accuracy under certain configurations in real-world IoMT applications.

Zhang et al. (2022) [20] introduced ACDP-Tree (Attribute Correlation-based Differential Privacy Tree), a privacy-preserving data publishing algorithm combining attribute correlation and differential privacy within a classification tree framework. Addressing the limitations of traditional k-

anonymity against consistency and background knowledge attacks, the method applied top-down iterative segmentation and Attribute Correlation Evaluation (ACE) to construct classification trees. The Adult dataset from UCI (University of California, Irvine), containing 48,842 records with 14 attributes, was used for experiments. Quasi-identifiers included age, marital-status, occupation, and sex, while capital-gain served as the sensitive attribute. Laplacian noise was added to ensure DP, with privacy budgets allocated per tree layer using class arithmetic methods. Results showed reduced absolute error, lower execution time, and minimal information loss, preserving data utility and privacy effectively. However, limitations involved the complexity of handling high-dimensional datasets and dependence on optimal parameter tuning for accurate performance across diverse medical data scenarios.

Vadrevu et al. (2020) [21] introduced two algorithms GenDP and Cluster-basedDP to enhance ε-Differential Privacy (DP) for relational and set-valued health data in privacy-preserving data publication. GenDP applied Personalized Differential Privacy (PDP) by generalizing data as homogeneous or heterogeneous using partitioning and classifiers, while Cluster-basedDP grouped data by similarity to handle multi-dimensional attributes. Both methods aimed to protect sensitive information without compromising utility. The algorithms were validated using a COVID-19 health dataset, and results demonstrated improved scalability and classification accuracy compared to existing DP models. The ε-DP framework ensured robust protection regardless of adversary background knowledge. The experimental evaluation showed that both models effectively balanced privacy and data utility across varying budget values. However, limitations included the dependency on optimal clustering and classification mechanisms, and potential complexity in handling highly diverse datasets, which impact accuracy and generalization in different real-world healthcare scenarios. The summary of the recent works in the privacy preservation in healthcare using differential privacy is given in Table III.

TABLE III. SUMMARY OF THE RECENT WORKS IN PRIVACY PRESERVATION IN HEALTHCARE USING DP

Author (s)	Methodology	Advantages	Disadvantages
Alsenani et al. (2025) [18]	DP	• Effectively applied multiple DP noise mechanisms. • Demonstrated efficient performance with increased client numbers.	• Limitations included assumptions of honest participants, leaving the system vulnerable to adversarial clients submitting manipulated summaries. • Lack of adversarial resilience and hybrid privacy mechanisms.
Wang and Li (2023) [19]	MDLDP	• Enhanced local privacy with data utility. • Preserved data correlation.	• Lack of improved encoding methods and optimization of perturbation techniques. • Challenges affected the scalability and accuracy under certain configurations in real-world IoMT applications.
Zhang et al. (2022) [20]	ACDP-Tree	• Improved privacy over k-anonymity. • Delivered low execution time, reduced absolute error, and minimal information loss.	• Limitations involved the complexity of handling high-dimensional datasets. • Heavy reliance on optimal parameter tuning for accurate performance across diverse medical data scenarios.
Vadrevu et al. (2020) [21]	ε-Differential Privacy	• Effective multi-dimensional data handling. • Effectively balanced privacy and data utility across varying budget values	• Limitations included the dependency on optimal clustering and classification mechanisms.

D. Secure Multi Party Computation for Privacy Preservation in Healthcare

SMPC is a cryptographic approach that enables multiple parties to collaboratively compute a function over their individual inputs without disclosing those inputs to each other. The privacy of all the participants is maintained by sharing only the final result, without revealing any individual data. This is especially useful in privacy-sensitive collaborations, such as hospitals jointly analyzing patient data without sharing confidential records. SMPC relies on secure protocols that prevent data leakage during computation. However, its implementation is complex and resource-intensive, particularly with many participants, as it involves multiple communication rounds that increase latency and reduce processing efficiency.

Dong et al. (2021) [22] evaluated the feasibility of SMPC using Yao's garbled circuits to support cross-institutional collaboration in clinical settings. The protocols such as Private Set Intersection-High Utilizer identification (PSI-HU) and Private Set Intersection Comorbidity Index calculation (PSI-CI) were tested on large, realistically synthesized datasets. Cuckoo hashing was used to enhance computational efficiency, delivering clinically relevant outputs within minutes. These protocols were provably secure in a semi-honest adversarial model and eliminated the need for a trusted third-party intermediary. Results demonstrated improved scalability and faster performance compared to traditional privacy-preserving record linkage (PPRL) systems. However, while memory efficiency allowed for scalability, limitations included the lack of support for heterogeneous clinical data such as genomics or mobile health. Moreover, the protocols were evaluated only in controlled environments, and broader deployment requires regulatory trust and institutional willingness to adopt new cryptographic methods.

Li et al. (2020) [23] proposed Chain-PPFL, a privacy-preserving framework using a chained SMPC technique. The model incorporated a Single-Masking mechanism to protect exchanged data and a Chained-Communication method for serially passing masked information among participants. Extensive experiments were conducted using MNIST and CIFAR-100 datasets under IID and Non-IID settings across three models such as Convolutional Neural Network, Multi-Layer Perceptron, and Limited-memory Broyden Fletcher Goldfarb Shanno (L-BFGS). Results showed that the model maintained high model accuracy and convergence speed, closely matching FedAVG, while offering strong privacy guarantees equivalent to differential privacy. Compared to DP-based FL schemes, Chain-PPFL exhibited superior

performance by effectively neutralizing added noise. However, the framework introduced additional communication costs and assumed reliable communication environments, such as Edge-Cloud or Smart City settings. Limitations also included lack of support for decentralized FL scenarios and potential scalability issues in more complex or unstable network infrastructures.

Li et al. (2020) [24] proposed a privacy-preserving self-serviced medical diagnosis scheme based on SMC. The methodology involved patients encrypting their medical data and sending it to a hospital server, which computed similarity between patient data and disease trait vectors to recommend treatment. The scheme ensured data confidentiality through privacy-preserving access control and resisted known security threats. The model leveraged HE to offload diagnostic operations to the hospital while preserving privacy. Real-world experiments validated the scheme's feasibility and performance under large-scale disease databases and multiple concurrent requests. The similarity-based disease matching approach enabled accurate diagnosis and reduced hospital burden. However, limitations included scalability issues due to constraints on the number of simultaneous requests and reliance on a trusted third party. Despite these, the approach offered secure and efficient diagnosis, empowering patients with insights into their health while maintaining data privacy.

Kumar et al. (2020) [25] introduced a privacy-preserving self-care health management system using SMPC to protect sensitive patient data during remote diagnosis, especially relevant during the COVID-19 pandemic. Patients shared encrypted health data with hospital servers, which was matched with existing medical records using a smart disease index. The system employed the Paillier encryption algorithm for partial HE, allowing computations on encrypted data without revealing patient information. This approach ensured data confidentiality and mitigated various security threats. The model aimed to support elderly patients unable to visit hospitals by enabling secure online consultations. Although effective in preserving privacy and offering reliable diagnosis, the system exhibited a significant limitation slow computational performance due to operations on encrypted data rather than plaintext. The approach showed potential for extension to diagnostic centers, pharmacies, and other secure e-service platforms like e-voting and e-auctions. The summary of the recent works in the privacy preservation in healthcare using secure multi-party computation is given in Table IV.

TABLE IV. SUMMARY OF THE RECENT WORKS IN PRIVACY PRESERVATION IN HEALTHCARE USING SMPC

Author (s)	Methodology	Advantages	Disadvantages
Dong et al. (2021) [22]	SMPC	• Achieved fast and scalable privacy-preserving computation in clinical collaborations • No trusted third-party needed	• Lack of support for heterogeneous clinical data such as genomics or mobile health. • Protocols were evaluated only in controlled environments, and broader deployment requires regulatory trust and institutional willingness to adopt new cryptographic methods.
Li et al. (2020) [23]	Chain-PPFL	• High accuracy with strong privacy • Resilient to noise impact.	• Lack of support for decentralized FL scenarios and potential scalability issues in more complex or unstable network infrastructures.
Li et al. (2020) [24]	SMPC	• Privacy-preserving diagnosis. • Scalable real- world validation.	• Limitations included scalability issues due to constraints on the number of simultaneous requests and reliance on a trusted third party.
Kumar et al. (2020) [25]	SMPC	• Remote self-care support. • Cross-domain applicability	• Limitation included the slow computational performance due to operations on encrypted data rather than plaintext

III. RESEARCH GAP

Despite the rapid advancements in data science in healthcare, robust privacy preservation remains a significant challenge. Existing methods such as HE, FL, DP and SMPC still face various limitations that hinder large-scale, real-world applications. However, their limitations reveal a persistent research gap in developing scalable, efficient, and adaptable privacy-preserving frameworks suitable for the complexity and diversity of healthcare environments. HE-based methods, while cryptographically robust, suffer from high computational complexity, inefficient handling of medical image data, and incompatibility with dynamic, decentralized settings like hospital networks or telemedicine platforms [5, 6]. FL models offer a promising approach by allowing model training without sharing raw data, yet it faces the lack of support for heterogeneous environments, high communication costs, and limited resilience to malicious participants or tampered models [13, 16]. Additionally, the privacy preservation mechanisms used in some of the FL often struggle when dealing with more complex and diverse healthcare data types, such as genomic data, wearable health device outputs, or dynamic patient records [11].

Moreover, DP struggles to manage adversarial clients, especially when the participants lack complete trust [18]. In some of the DP approaches, the lack of optimization in perturbation techniques and optimal parameter tuning affects its utility and accuracy while applied to high dimensional or highly diverse clinical datasets [19]. SMPC protocols are currently constrained by issues of scalability and performance, especially in heterogeneous environments. It requires a trusted third party, which is a single point of failure, and their performance degrades due to slow computational processes when dealing with encrypted data. To address these limitations in privacy-preserving for health care, there is a need for a communication-efficient, scalable privacy-preserving model that integrates FL with advanced encryption techniques, ensuring robust data security while supporting diverse and high-dimensional medical data. Furthermore, the model must be designed to handle heterogeneous healthcare environments, offering interoperability across different systems and maintaining low-latency performance to meet the demands of real-time clinical applications.

IV. CONCLUSION

The study highlights the growing importance of privacy-preserving data science in healthcare, where the protection of

sensitive patient information is as critical as data-driven innovation. Through a systematic survey of Federated Learning, Homomorphic Encryption, Differential Privacy, and Secure Multi-Party Computation, this paper emphasizes how each method contributes to safeguarding confidentiality while enabling efficient data analytics. Federated Learning ensures secure model training without centralizing data, while Homomorphic Encryption allows computations on encrypted data, maintaining end-to-end confidentiality. Differential Privacy protects individual identities by adding calibrated noise to results, and SMPC enables collaborative computation without direct data sharing. Despite these advancements, challenges such as computational overhead, limited scalability, and lack of interoperability across diverse healthcare systems remain unresolved. The analysis underscores the need for hybrid frameworks that integrate FL with lightweight encryption and adaptive privacy budgets to balance performance with protection. Future research should focus on quantum-resilient cryptography, efficient key management, and interoperability standards to support large-scale, real-time healthcare analytics. Additionally, ethical compliance and regulatory alignment with frameworks like HIPAA and GDPR must be embedded within algorithmic design to foster trust and accountability. Overall, achieving secure and privacy-preserving healthcare data science requires convergence of cryptographic innovation, distributed intelligence, and governance, enabling a trustworthy and sustainable digital healthcare ecosystem.

REFERENCES

[1] Sahi, M. A., Abbas, H., Saleem, K., Yang, X., Derhab, A., Orgun, M. A., ... & Yaseen, A. (2017). Privacy preservation in e-healthcare environments: State of the art and future directions. Ieee Access, 6, 464-478.

[2] Singh, S., Rathore, S., Alfarraj, O., Tolba, A., & Yoon, B. (2022). A framework for privacy-preservation of IoT healthcare data using Federated Learning and blockchain technology. *Future Generation Computer Systems*, *129*, 380-388.

[3] Jin, H., Luo, Y., Li, P., & Mathew, J. (2019). A review of secure and privacy-preserving medical data sharing. IEEE access, 7, 61656-61669.

[4] Gandhi, B. M., Vaghadia, S. B., Kumhar, M., Gupta, R., Jadav, N. K., Bhatia, J., ... & Alabdulatif, A. (2025). Homomorphic encryption and collaborative machine learning for secure healthcare analytics. *Security and Privacy*, *8*(1), e460.

[5] Naresh, V. S., & Reddi, S. (2025). Exploring the future of privacy-preserving heart disease prediction: a fully homomorphic encryption-driven logistic regression approach. Journal of Big Data, 12(1), 52.

[6] Babu, K. M., Syed, M., Shaik, S., Thalari, S., Macha, U., & Chatakondu, A. (2025). Fully homomorphic encryption framework for privacy preserving in healthcare through decentralized machine learning. In Challenges in Information, Communication and Computing Technology (pp. 812-816). CRC Press.

[7] Panzade, P., Takabi, D., & Cai, Z. (2024). MedBlindTuner: Towards Privacy-Preserving Fine-Tuning on Biomedical Images with Transformers and Fully Homomorphic Encryption. In AI for Health Equity and Fairness: Leveraging AI to Address Social Determinants of Health (pp. 197-208). Cham: Springer Nature Switzerland.

[8] Sarkar, E., Chielle, E., Gursoy, G., Chen, L., Gerstein, M., & Maniatakos, M. (2023). Privacy-preserving cancer type prediction with homomorphic encryption. Scientific reports, 13(1), 1661.

[9] Ali, A., Al-Rimy, B. A. S., Alsubaei, F. S., Almazroi, A. A., & Almazroi, A. A. (2023). Healthlock: Blockchain-based privacy preservation using homomorphic encryption in internet of things healthcare applications. Sensors, 23(15), 6762.

[10] Darzidehkalani, E., Ghasemi-Rad, M., & Van Ooijen, P. M. A. (2022). Federated learning in medical imaging: part I: toward multicentral health care ecosystems. *Journal of the american college of radiology*, *19*(8), 969-974.

[11] Meduri, K., Nadella, G. S., Yadulla, A. R., Kasula, V. K., Maturi, M. H., Brown, S., ... & Gonaygunta, H. (2025). Leveraging federated learning for privacy-preserving analysis of multi-institutional electronic health records in rare disease research. Journal of Economy and Technology, 3, 177-189.

[12] Yang, M., Huang, D., Wan, W., & Jin, M. (2024). Federated learning for privacy-preserving medical data sharing in drug development. Applied and Computational Engineering, 108, 7-13.

[13] Muthalakshmi, M., Jeyapal, K., Vinoth, M., PS, D., Murugan, N. S., & Sheela, K. S. (2024, August). Federated Learning for Secure and Privacy-Preserving Medical Image Analysis in Decentralized Healthcare Systems. In 2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 1442-1447). IEEE.

[14] Markkandan, S., Bhavani, N. P. G., & Nath, S. S. (2024). A privacy-preserving expert system for collaborative medical diagnosis across multiple institutions using federated learning. Scientific Reports, 14(1), 22354.

[15] Abaoud, M., Almuqrin, M. A., & Khan, M. F. (2023). Advancing federated learning through novel mechanism for privacy preservation in healthcare applications. IEEE Access, 11, 83562-83579.

[16] Zhang, L., Xu, J., Vijayakumar, P., Sharma, P. K., & Ghosh, U. (2022). Homomorphic encryption-based privacy-preserving federated learning in IoT-enabled healthcare system. IEEE Transactions on Network Science and Engineering, 10(5), 2864-2880.

[17] Li, J., Meng, Y., Ma, L., Du, S., Zhu, H., Pei, Q., & Shen, X. (2021). A federated learning-based privacy-preserving smart healthcare system. IEEE Transactions on Industrial Informatics, 18(3).

[18] Alsenani, Y. (2025). FAItH: Federated Analytics and Integrated Differential Privacy with Clustering for H ealthcare Monitoring. Scientific Reports, 15(1), 10155.

[19] Wang, J., & Li, X. (2023). Secure medical data collection in the internet of medical things based on local differential privacy. Electronics, 12(2), 307.

[20] Zhang, S., & Li, X. (2022). Differential privacy medical data publishing method based on attribute correlation. Scientific Reports, 12(1), 15725.

[21] Vadrevu, P. K., Adusumalli, S. K., & Mangalapalli, V. K. (2020). A hybrid approach for personal differential privacy preservation in homogeneous and heterogeneous health data sharing. High Technology Letters, 26(9), 1223-1239.

[22] Dong, X., Randolph, D. A., Weng, C., Kho, A. N., Rogers, J. M., & Wang, X. (2021). Developing high performance secure multi-party computation protocols in healthcare: a case study of patient risk stratification. AMIA Summits on Translational Science Proceedings, 2021, 200.

[23] Li, Y., Zhou, Y., Jolfaei, A., Yu, D., Xu, G., & Zheng, X. (2020). Privacy-preserving federated learning framework based on chained secure multiparty computing. IEEE Internet of Things Journal, 8(8), 6178-6186.

[24] Li, D., Liao, X., Xiang, T., Wu, J., & Le, J. (2020). Privacy-preserving self-serviced medical diagnosis scheme based on secure multi-party computation. Computers & Security, 90, 101701.

[25] Kumar, A. V., Sujith, M. S., Sai, K. T., Rajesh, G., & Yashwanth, D. J. S. (2020, December). Secure Multiparty computation enabled E-Healthcare system with Homomorphic encryption. In IOP Conference Series: Materials Science and Engineering (Vol. 981, No. 2, p. 022079). IOP Publishing.