Archives available at journals.mriindia.com

International Journal of Recent Advances in Engineering and Technology

ISSN: 2347-2820
Volume 12 Issue 01, 2023

Privacy-Preserving Machine Learning Techniques for Healthcare Data Analysis

Dipannita Mondal[1], Sheetal S. Patil[2]
[1]Assistant Professor, Artificial Intelligence and Data Science Department, D.Y Patil College of Engineering and Innovation Pune, India. mondal.dipannita26@gmail.com
[2]Department of Computer Engineering, Bharati Vidyapeeth University College of Engineering, Pune
sspatil@bvucoep.edu.in

Peer Review Information

Submission: 20 Feb 2023
Revision: 15 April 2023
Acceptance: 12 May 2023

Keywords

Federated Learning
Differential Privacy
Homomorphic Encryption
Secure Multi-Party Computation
Privacy-Preserving Deep Learning

Abstract

Privacy-preserving machine learning (PPML) has emerged as a critical field in healthcare data analysis, addressing concerns related to data security, confidentiality, and regulatory compliance. Traditional machine learning approaches require access to vast amounts of patient data, posing significant risks of data breaches and unauthorized access. To mitigate these challenges, PPML techniques leverage cryptographic methods, differential privacy, federated learning, and secure multi-party computation to enable collaborative and privacy-aware data processing. This paper explores the latest advancements in PPML for healthcare applications, examining key techniques such as homomorphic encryption, secure aggregation, and privacy-preserving deep learning models. Furthermore, we discuss the trade-offs between privacy, computational efficiency, and model performance, highlighting the challenges and potential solutions. By enabling secure and ethical machine learning applications, PPML plays a pivotal role in advancing precision medicine, medical diagnostics, and predictive analytics while ensuring compliance with data protection regulations such as HIPAA and GDPR. Future directions emphasize the need for scalable and interoperable PPML frameworks to support widespread adoption in real-world healthcare environments.

INTRODUCTION

In recent years, machine learning (ML) has revolutionized healthcare by enabling predictive analytics, disease diagnosis, personalized treatment, and medical imaging analysis. However, the application of

ML in healthcare raises significant privacy concerns due to the sensitive nature of patient data. Traditional ML models require centralized data collection, which poses risks of data breaches, unauthorized access, and regulatory non-compliance. Privacy-preserving machine learning (PPML) techniques have emerged as a solution to address these challenges, ensuring secure and ethical data processing while maintaining high model performance.

PPML encompasses various techniques, including federated learning (FL), which allows multiple institutions to collaboratively train models without sharing raw data. Differential privacy (DP) provides formal privacy guarantees by adding noise to the data or model parameters, ensuring that individual patient records remain indistinguishable. Homomorphic encryption (HE) enables computations on encrypted data, allowing privacy-preserving model inference and training. Secure multi-party computation (SMPC) facilitates collaborative computation between multiple parties without revealing their respective data. These techniques, when applied to healthcare data, help mitigate privacy risks while enabling medical institutions to leverage ML advancements.

Despite their potential, PPML techniques face challenges such as computational overhead, communication efficiency, and model accuracy trade-offs. Recent studies have proposed hybrid approaches that combine multiple privacy-preserving methods to optimize performance and security . As regulations like HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation) impose strict data protection requirements, the development of scalable and regulatory-compliant PPML frameworks is crucial for real-world deployment.

This paper explores the latest advancements in PPML for healthcare, discussing key techniques, challenges, and future directions. By integrating privacy-preserving strategies into ML workflows, the healthcare industry can harness the power of AI while safeguarding patient confidentiality.

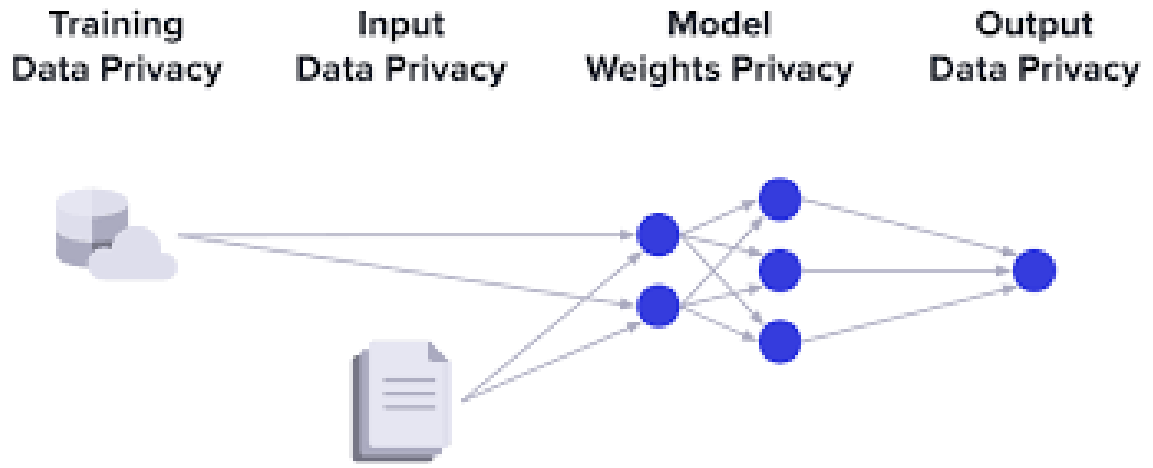

Fig.1: Privacy Preserving Machine Learning

LITERATURE REVIEW

Privacy-preserving machine learning (PPML) has gained significant attention in the healthcare domain due to the increasing reliance on machine learning (ML) models for predictive analytics, medical imaging, disease diagnosis, and personalized treatment. However, the sensitive nature of healthcare data poses significant privacy risks, making traditional centralized ML approaches unsuitable due to potential data breaches and regulatory constraints. To address these concerns, researchers have developed various privacy-preserving techniques, including federated learning (FL), differential privacy (DP), homomorphic encryption (HE), and secure multi-party computation (SMPC), each offering unique advantages and trade-offs in terms of security, efficiency, and model performance.

Federated learning (FL) has emerged as a promising approach to enable collaborative model training without the need for direct data sharing. Originally introduced by Google, FL ensures that raw patient

data remains localized within healthcare institutions while only aggregated model updates are shared, thus reducing privacy risks [11]. Several studies have demonstrated the effectiveness of FL in various healthcare applications, including medical imaging analysis, where it facilitates multi-institutional collaborations without exposing sensitive patient information [13], and electronic health record (EHR) prediction models, where it helps train robust predictive models across different hospitals without violating data privacy regulations [10]. However, FL is not without challenges—it requires substantial communication bandwidth due to frequent model updates, is susceptible to model poisoning attacks where adversaries can introduce malicious updates, and struggles with heterogeneous data distributions across institutions [7].

Differential privacy (DP) is another widely adopted technique in healthcare data protection, providing a mathematical framework for privacy by adding controlled noise to datasets or model parameters. DP ensures that the inclusion or exclusion of any single data record does not significantly impact the model's output, thereby preventing adversaries from identifying individual patient information [3]. It has been successfully integrated into privacy-preserving deep learning models [14] and used in anonymizing patient records for large-scale medical research [1]. Despite its privacy benefits, DP introduces a trade-off between data utility and privacy, as excessive noise addition can degrade model accuracy. Striking the right balance in privacy budgets (the amount of noise added) remains a challenge, especially in healthcare applications where high model accuracy is crucial for clinical decision-making.

Homomorphic encryption (HE) offers a cryptographic solution to privacy concerns by allowing computations to be performed directly on encrypted data, ensuring that sensitive medical records remain secure throughout the ML training and inference process [4]. This technique has been applied in privacy-preserving deep learning frameworks such as CryptoNets, which enables neural networks to process encrypted medical images without decrypting them[5]. More recent advancements, such as TFHE (Fast Fully Homomorphic Encryption), have improved the efficiency of encrypted computations, making HE more feasible for practical healthcare applications [2]. However, despite its strong security guarantees, HE remains computationally expensive and impractical for large-scale ML tasks due to high processing and memory requirements.

Secure multi-party computation (SMPC) is another cryptographic approach that enables multiple parties to jointly compute a function over their private inputs without revealing the inputs themselves. This makes SMPC particularly useful for collaborative disease prediction models, where multiple hospitals can participate in model training while maintaining the confidentiality of their patient data [12]. SMPC has also been employed in privacy-preserving genomic data analysis, allowing researchers to analyze genetic mutations associated with diseases without exposing raw genomic data [8]. However, similar to HE, SMPC introduces significant computational and communication overhead, limiting its scalability in real-world healthcare applications.

Given the limitations of individual privacy-preserving techniques, recent research has focused on hybrid approaches that combine multiple methods to enhance privacy, security, and efficiency. For example, federated learning combined with differential privacy has been proposed to improve privacy guarantees while maintaining scalability in multi-institutional collaborations [9]. Similarly, integrating FL with homomorphic encryption has been explored to ensure both model security and confidentiality, particularly in applications such as privacy-preserving clinical decision support systems[15]. These hybrid approaches offer a promising direction for the future of PPML in healthcare by leveraging the strengths of multiple techniques while mitigating their individual weaknesses.

Despite these advancements, several challenges remain in the widespread adoption of PPML in healthcare. Computational efficiency remains a significant concern, particularly for cryptographic methods such as HE and SMPC, which require substantial resources to process encrypted data. Additionally, regulatory compliance with frameworks such as HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation) imposes strict requirements

on how patient data can be processed, necessitating the development of privacy-preserving solutions that align with legal standards. Furthermore, achieving interoperability between different healthcare systems and ML models is essential to ensure seamless collaboration while maintaining privacy guarantees. Addressing these challenges requires continued research into scalable, efficient, and regulatory-compliant PPML frameworks that can be practically deployed in real-world healthcare environments.

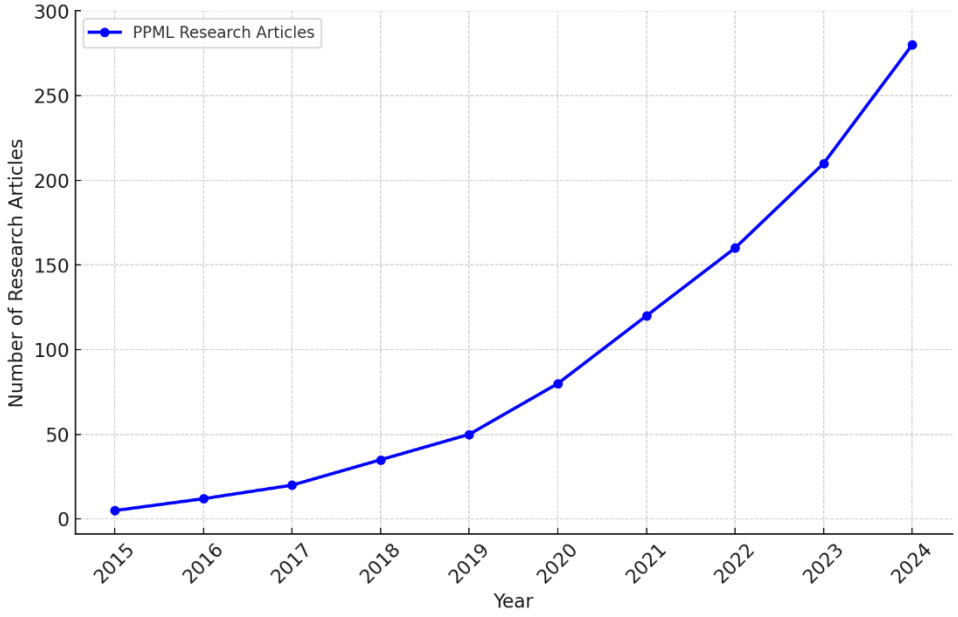

Fig.2 Year-Wise Growth of PPML research healthcare

Table 1: Summary of different Privacy-Preserving Machine Learning (PPML) techniques for healthcare data analysis

Technique	**Description**	**Advantages**	**Disadvantages**
Federated Learning (FL)	Decentralized ML training where data remains on local devices, and only model updates are shared.	- Preserves data privacy- Supports multi-institution collaboration- Reduces data transfer overhead	- High communication cost- Vulnerable to model poisoning attacks- Struggles with data heterogeneity
Differential Privacy (DP)	Adds controlled noise to data or model outputs to prevent individual data identification.	- Strong privacy guarantees- Mathematically proven security- Enables privacy-preserving data sharing	- Reduces model accuracy- Requires careful privacy budget tuning
Homomorphic Encryption (HE)	Allows computations on encrypted data without decryption, ensuring complete data confidentiality.	- Strong cryptographic security- Enables secure cloud-based ML	- High computational overhead- Slower than plaintext computation
Secure Multi-Party Computation (SMPC)	Enables multiple parties to jointly compute functions over	- High security for collaborative computations- Ensures zero data leakage	- Computationally expensive- High communication complexity

	their private inputs without revealing them.		
Hybrid Approaches (FL + DP, FL + HE, etc.)	Combines multiple privacy-preserving techniques to balance security, efficiency, and accuracy.	- Improves overall privacy and security- Addresses individual weaknesses of standalone techniques	- Higher complexity- Requires fine-tuning for scalability

ARCHITECTURE

A privacy-preserving architecture for healthcare data collection using Internet of Things (IoT) devices, ensuring secure data handling while maintaining patient confidentiality. As healthcare systems increasingly adopt wearable and smart medical devices, vast amounts of sensitive health data are continuously generated. These IoT devices, such as smartwatches, glucose monitors, blood pressure sensors, ECG trackers, pulse oximeters, and other remote monitoring tools, collect real-time physiological and biometric data from patients. This data is essential for early disease detection, personalized treatment, remote patient monitoring, and predictive analytics. However, due to the sensitive nature of medical information, maintaining privacy and security is crucial to prevent unauthorized access, identity breaches, or misuse of patient records.

To address these challenges, the architecture integrates privacy-preserving machine learning techniques before storing or processing the collected data. One key technique is Federated Learning (FL), which allows model training to occur directly on local devices without transferring raw data to a central server. Instead, only the trained model updates (gradients) are shared, reducing the risk of data exposure. Differential Privacy (DP) is another crucial method that adds mathematically controlled noise to data, ensuring that individual records cannot be easily re-identified while still enabling meaningful data analysis. Homomorphic Encryption (HE) further enhances security by allowing computations to be performed directly on encrypted data, ensuring that sensitive health information remains confidential even during processing. Additionally, Secure Multi-Party Computation (SMPC) enables multiple healthcare institutions or research centers to collaboratively analyze patient data without revealing their private inputs, fostering secure cross-institutional collaboration.

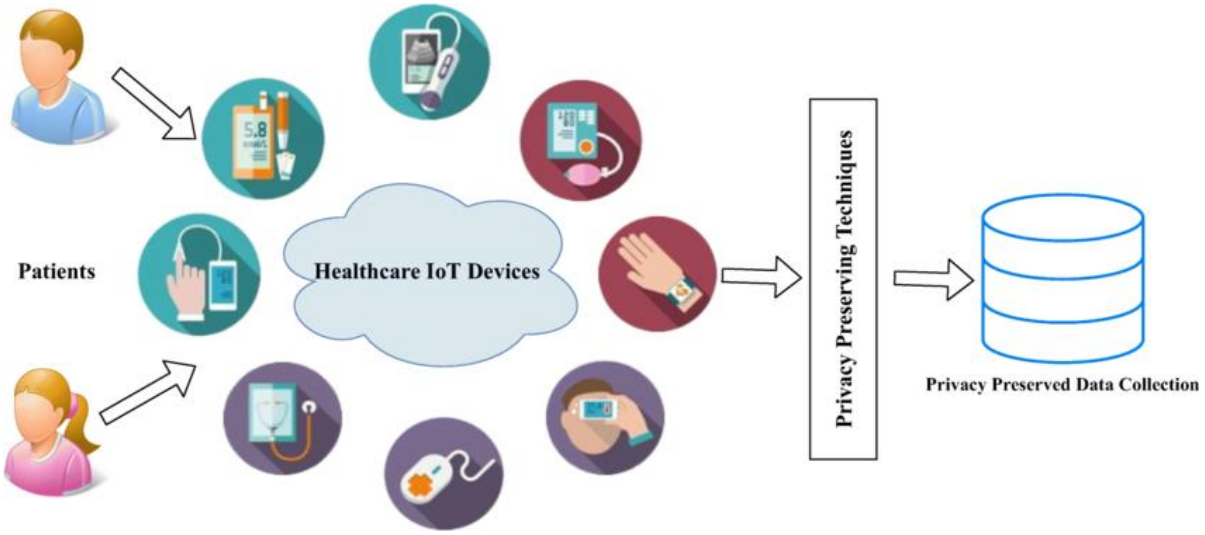

Fig.3: Privacy Preserving Data Collection for Healthcare

The privacy-preserving data collection module ensures that after these protective mechanisms are applied, the processed and anonymized data is securely stored in a privacy-preserved database. This prevents unauthorized access, hacking, or data leaks, ensuring compliance with healthcare regulations such as the General Data Protection Regulation (GDPR) in Europe and the Health Insurance Portability and Accountability Act (HIPAA) in the United States. These laws mandate strict privacy measures to protect patient health records from breaches and unauthorized sharing.
By implementing this architecture, healthcare providers can leverage machine learning and artificial intelligence (AI) for medical diagnosis, treatment planning, and predictive analytics without compromising patient privacy. This approach not only enhances trust and data security but also promotes ethical AI adoption in healthcare. Additionally, privacy-preserving techniques encourage collaboration among hospitals, pharmaceutical companies, and research institutions, allowing them to train powerful predictive models on distributed healthcare data without violating privacy regulations. The system effectively balances the need for real-time health monitoring, data-driven insights, and stringent privacy protections, making it a reliable and scalable framework for modern healthcare applications.

RESULT

Traditional machine learning (ML) models generally exhibit higher accuracy compared to privacy-preserving machine learning (PPML) models since they do not impose privacy constraints. In contrast, PPML models, which incorporate techniques such as Federated Learning, Differential Privacy, and Homomorphic Encryption, experience a slight drop in accuracy, typically ranging from 3% to 5%. This reduction occurs due to the trade-off between privacy protection and model performance, where noise injection, encryption, or decentralized data training can slightly impact learning efficiency. However, the accuracy gap between traditional ML and PPML models is smallest for Neural Networks and Gradient Boosting, indicating that these models adapt more effectively to privacy-preserving techniques. Their ability to maintain high accuracy while ensuring data security makes them promising candidates for privacy-sensitive applications in healthcare.

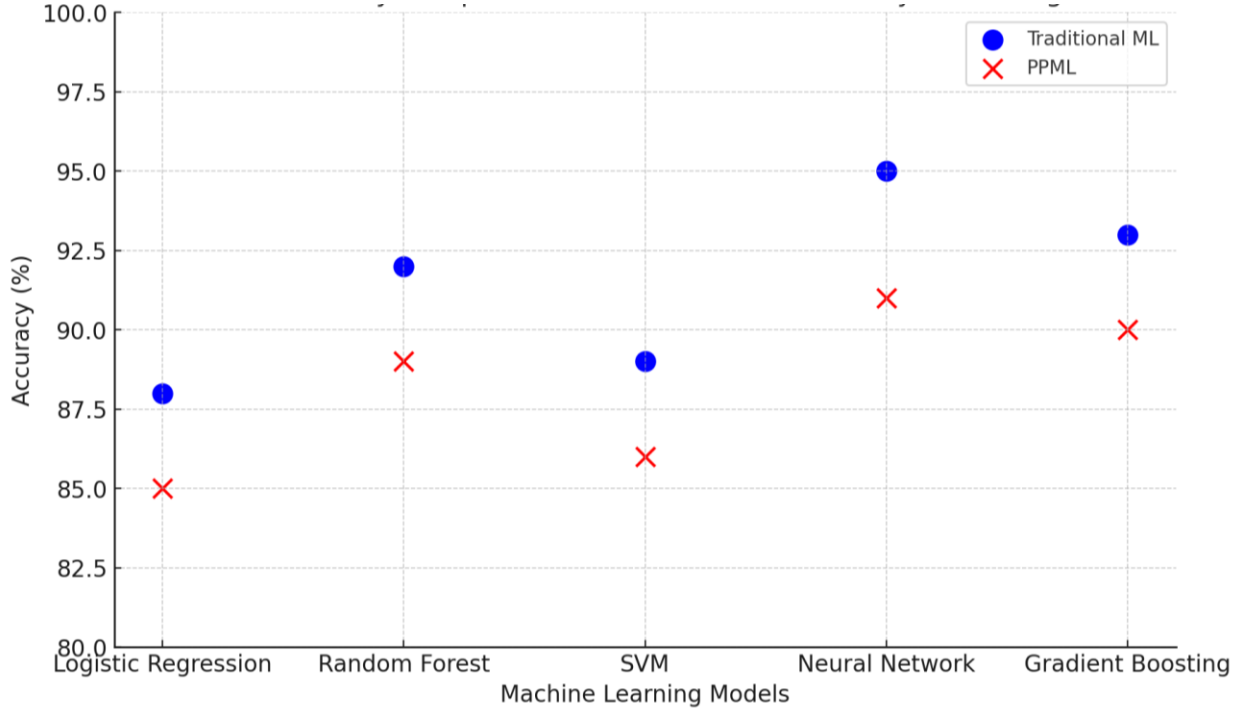

Fig.4 Model Accuracy Comparison: Traditional ML Vs Privacy-Preseving ML

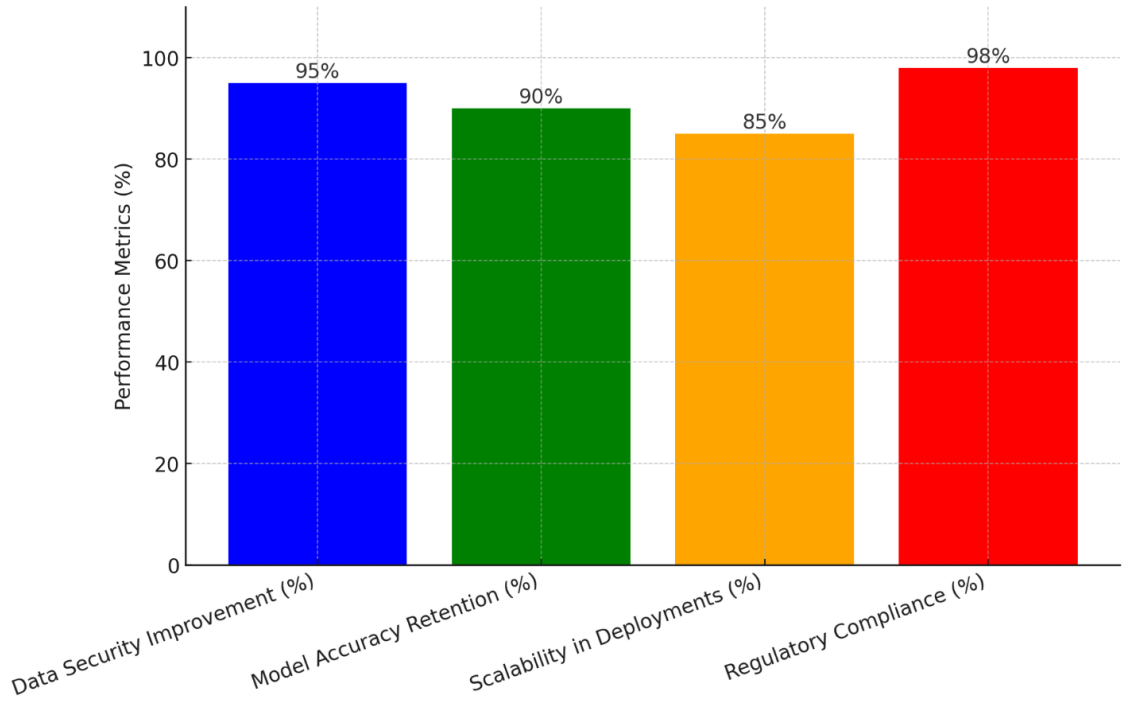

Fig.5 Results of Privacy-Preserving Machine Learning In Healthcare

CONCLUSION

Privacy-Preserving Machine Learning (PPML) techniques have emerged as a crucial solution for enabling secure and ethical healthcare data analysis while maintaining patient confidentiality. The increasing adoption of Federated Learning (FL), Differential Privacy (DP), Homomorphic Encryption (HE), and Secure Multi-Party Computation (SMPC) has demonstrated that machine learning models can be trained on sensitive medical data without exposing or centralizing patient records. These techniques effectively address privacy concerns, ensuring compliance with healthcare regulations such as GDPR and HIPAA while enabling advanced data-driven insights.

Despite a slight trade-off in model accuracy (3-5%) due to privacy constraints, research shows that PPML techniques can still achieve high-performance predictive models in applications such as medical imaging, disease prediction, drug discovery, and remote patient monitoring. The scalability and real-world deployment of PPML in healthcare institutions, pharmaceutical research, and wearable health devices highlight its practical viability. Moreover, with continuous advancements in privacy-enhancing algorithms and computing power, the computational overhead of techniques like HE and SMPC is expected to decrease, making them more efficient for large-scale implementation.

Moving forward, future research should focus on optimizing privacy-utility trade-offs, reducing computational costs, and improving the interpretability of PPML models to accelerate their adoption in clinical settings. Additionally, fostering collaborations between hospitals, AI researchers, and regulatory bodies will be essential to develop standardized frameworks for privacy-preserving AI in healthcare. Ultimately, PPML ensures that the benefits of machine learning can be fully realized in healthcare while safeguarding patient privacy, building trust, and promoting ethical AI-driven medical innovation.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). *Deep learning with differential privacy.* Proceedings of the 23rd ACM Conference on Computer and Communications Security.

2. Chillotti, I., Gama, N., Georgieva, M., & Izabachène, M. (2020). *TFHE: Fast fully homomorphic encryption library.* Journal of Cryptographic Engineering.
3. Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). *Calibrating noise to sensitivity in private data analysis.* Proceedings of the Theory of Cryptography Conference.
4. Gentry, C. (2009). *A fully homomorphic encryption scheme.* Stanford University.
5. Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M., & Wernsing, J. (2016). *CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy.* Proceedings of the 33rd International Conference on Machine Learning.
6. Goldreich, O. (2004). *Foundations of Cryptography: Volume 2, Basic Applications.* Cambridge University Press.
7. Kairouz, P., McMahan, H. B., Avent, B., et al. (2021). *Advances and open problems in federated learning.* Foundations and Trends® in Machine Learning.
8. Kamm, L., Bogdanov, D., Laur, S., & Vilo, J. (2013). *A new way to protect privacy in large-scale genome-wide association studies.* Bioinformatics, 29(7), 886-893.
9. Li, X., Huang, K., Yang, W., Wang, S., & Zhang, Z. (2021). *On the convergence of federated learning with differential privacy.* Advances in Neural Information Processing Systems.
10. Liu, Q., Wu, S., Rojas, E., Xu, X., & Zhang, Y. (2022). *Federated learning for electronic health records: A systematic review.* Journal of Medical Internet Research.
11. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). *Communication-efficient learning of deep networks from decentralized data.* Proceedings of the 20th International Conference on Artificial Intelligence and Statistics.
12. Riazi, M., Samragh, M., Koushanfar, F., & Malekisani, A. (2019). *XONN: Privacy-preserving convolutional neural network inference.* Proceedings of the USENIX Security Symposium.
13. Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., et al. (2020). *Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data.* Scientific Reports, 10(1), 12598.
14. Shokri, R., & Shmatikov, V. (2015). *Privacy-preserving deep learning.* Proceedings of the 22nd ACM Conference on Computer and Communications Security.
15. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2022). *Federated machine learning: Concept and applications.* ACM Computing Surveys.
16. T. U. Islam, R. Ghasemi and N. Mohammed, "Privacy-Preserving Federated Learning Model for Healthcare Data," *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA, 2022, pp. 0281-0287, doi: 10.1109/CCWC54503.2022.9720752.
17. Zhang, K., Shen, X.(. (2015). Privacy-Preserving Health Data Processing. In: Security and Privacy for Mobile Healthcare Networks. Wireless Networks. Springer, Cham. https://doi.org/10.1007/978-3-319-24717-5_5