IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2025

DOI: 10.17148/IJARCCE.2025.14643

A Survey on Privacy-Preserving Data Imputation via Multi-Party Computation for Medical Applications

Shruthi T S[1], Raghusai Achuth[2], Manoja G V[3], Pervez Ansari[4], Syed Farhan[5]

Assistant Professor, Dept. of CSE, KSIT, Bengaluru, India[1]

Student, Dept. of CSE, KSIT, Bengaluru, India[2]

Student, Dept. of CSE, KSIT, Bengaluru, India[3]

Student, Dept. of CSE, KSIT, Bengaluru, India[4]

Student, Dept. of CSE, KSIT, Bengaluru, India[5]

Abstract: Medical datasets frequently contain missing values, which can negatively impact machine learning models used in healthcare. However, imputing these values while ensuring patient privacy presents a significant challenge. This survey explores various privacy-preserving data imputation techniques, with a focus on Secure Multi-Party Computation (MPC). We review four imputation methods—mean, median, regression, and k-nearest neighbors (KNN)—and how each can be implemented securely in distributed medical environments. The paper also discusses hybrid approaches, integration with differential privacy, and federated settings. Our analysis concludes that MPC-based imputation provides strong privacy guarantees with high accuracy, paving the way for privacy-conscious medical data analysis.

Keywords: Data Imputation, Medical Data Privacy, Multi-Party Computation (MPC), Secure Computation, Privacy-Preserving Machine Learning

I. INTRODUCTION

Healthcare analytics increasingly depends on machine learning models to deliver predictions, diagnoses, and treatment suggestions. However, missing data—due to sensor failure, entry errors, or incomplete patient records—can compromise data integrity. Traditional imputation strategies violate confidentiality.

With regulations such as HIPAA and GDPR imposing strict data-sharing constraints, privacy-preserving computation is critical. Secure Multi-Party Computation (MPC) offers a cryptographic solution allowing multiple institutions to collaborate on data analysis without disclosing individual records. This paper surveys the use of MPC in medical data imputation, reviewing techniques, tools, and challenges.

II. LITERATURE SURVEY

The issue of missing data in healthcare is long-standing, and privacy-preserving imputation has recently emerged as a critical focus of research due to increasing regulatory concerns (e.g., HIPAA, GDPR) and data-sharing needs across institutions. The literature reflects efforts spanning secure computation, differential privacy, federated learning, and hybrid approaches.

[1] A. Traditional Imputation Techniques and Limitations

Conventional techniques for handling missing data include listwise deletion, mean/mode substitution, regression imputation, and k-nearest neighbors (KNN). These techniques are computationally inexpensive and widely adopted but assume access to complete data. As a result, they are unsuitable when data is distributed across multiple private sources such as hospitals. Furthermore, simple approaches like mean substitution introduce statistical bias, while more complex methods like KNN suffer from scalability issues.

[2] B. Secure Multi-Party Computation (MPC) in Imputation

Secure Multi-Party Computation (MPC) enables joint computations over private datasets without exposing individual data points. Jentsch et al. (2024) developed MPC-based implementations of mean, median, regression, and KNN imputation and tested them on medical datasets. Their results demonstrated that MPC can achieve near-plaintext accuracy

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2025
DOI: 10.17148/IJARCCE.2025.14643

with strong privacy guarantees. However, the computation time, especially for KNN, increases significantly with data size due to secure distance calculations and sorting mechanisms.

Frameworks such as MP-SPDZ and CECILIA have enabled the implementation of arithmetic and linear algebra operations within MPC settings, including secure LU decomposition used in regression imputation. These systems use secret sharing or homomorphic encryption as the underlying cryptographic primitive.

[3] C. Differential Privacy (DP) for Imputation

Differential Privacy (DP) introduces random noise into computations to prevent inference attacks, offering a statistical privacy guarantee. Das et al. (2022) explored DP-based imputation using Laplace and Gaussian noise in statistical estimates. While DP methods are scalable and fast, they often reduce imputation accuracy—especially for high-dimensional or sparse datasets. This creates a trade-off between privacy budget (ε) and utility. Unlike MPC, DP does not require multiple parties or communication overhead but cannot achieve perfect reconstruction of original statistics.

[4] D. Federated Learning and Hybrid Techniques

Federated Learning (FL) offers a collaborative framework where multiple parties train a shared model without transferring data. However, preprocessing steps like imputation must also respect data locality and privacy. While FL itself doesn't solve missing data issues, it has encouraged the development of federated imputation frameworks.
Hybrid approaches that combine MPC and DP have been proposed to achieve scalability and strong privacy. For example, some systems use MPC for high-sensitivity features and DP for low-sensitivity ones. This strategy ensures that the privacy-utility trade-off is balanced. Adaptive policy engines are used to dynamically decide which technique to apply per feature based on sensitivity and missingness.

[5] E. Homomorphic Encryption and Secure Aggregation

Gürsoy et al. (2022) introduced privacy-preserving genotype imputation using Fully Homomorphic Encryption (FHE), which allows computation over encrypted data. Though FHE ensures very strong privacy, it is prohibitively slow for real-time or large-scale applications. Secure aggregation techniques such as those used in federated analytics offer better performance but are limited in operation types.

[6] F. Medical Applications and Practical Considerations

Liao et al. (2014) highlighted that imputation methods must be chosen carefully for biomedical data due to its sparsity, high dimensionality, and heterogeneity. They emphasized that regression-based techniques are more reliable when inter-feature correlation is strong. Additionally, Jagannathan and Wright (2008) developed early models of privacy-preserving data mining and imputation for clinical datasets, though they lacked scalability for modern EHR systems.

In practical deployment, factors like semi-honest vs malicious adversary models, computation time, communication overhead, and interoperability with existing hospital IT systems remain barriers.

III. OBJECTIVES

- **To conduct a comprehensive survey of imputation techniques implemented using Secure Multi-Party Computation (MPC)**, specifically tailored for handling missing values in sensitive medical datasets distributed across multiple parties.
- **To evaluate the computational scalability and performance efficiency** of various MPC-based imputation algorithms (such as mean, median, regression, and k-nearest neighbors) under realistic healthcare data constraints.
- **To critically examine the trade-offs between privacy preservation and imputation accuracy**, with a focus on balancing data utility and compliance with privacy regulations such as HIPAA and GDPR.
- **To identify current research limitations and emerging challenges** in privacy-preserving imputation, and to propose future research directions for practical deployment in federated healthcare environments.

IV. METHODOLOGY

This section presents the proposed methodology for privacy-preserving data imputation using Secure Multi-Party Computation (MPC). The framework is designed for medical datasets that are vertically partitioned across multiple institutions. It ensures that sensitive patient information is never exposed during computation, addressing privacy concerns while enabling accurate imputation.

A. Overall Framework

In the proposed system, each participating institution retains its data locally while participating in a secure collaborative imputation process. The MPC protocol is used to securely compute missing values without revealing any private input data. The key steps involved are:

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2025
DOI: 10.17148/IJARCCE.2025.14643

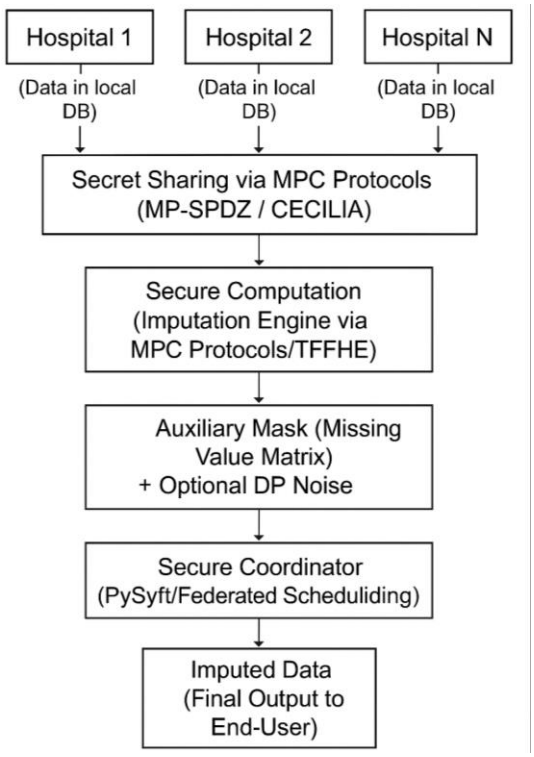
1. **Data Collection and Secret Sharing**: Each hospital or healthcare entity divides its data into secret shares using additive or Shamir secret-sharing schemes. These shares are distributed to computing nodes for secure processing.
2. **Missing Value Identification**: An auxiliary binary matrix is generated to indicate the presence (1) or absence (0) of each value in the dataset. This mask guides the imputation logic.
3. **Secure Imputation Computation**: Based on the mask and data type, an appropriate imputation strategy is applied. All operations are executed securely over shared data using MPC protocols.
4. **Result Reconstruction**: After imputation, the secret shares of the computed values are combined to reveal the final dataset with missing values filled, preserving both privacy and accuracy.

B. Secure Implementation of Imputation Algorithms

The following imputation methods have been securely implemented under MPC protocols:

- **Mean/Mode Imputation**: For numerical features, the secure sum and count of non-missing values are computed, followed by fixed-point division. For categorical features, secure counting is used to determine the mode.
- **Median Imputation**: Secure sorting protocols (e.g., oblivious sorting networks) are used to find the median. This is particularly useful for skewed data distributions.
- **Regression Imputation**: A secure version of multivariate linear regression is implemented using LU decomposition and matrix multiplication protocols. It is applied when there is a strong correlation between features.
- **k-Nearest Neighbors (KNN) Imputation**: Pairwise distances between data points are calculated using secure arithmetic, followed by secure top-k selection and aggregation. While accurate, this method is computationally intensive.

C. System Architecture and Tools



The system operates in a **federated privacy-preserving architecture**, Each node represents a hospital or data source, contributing to the computation without sharing raw data. A central secure coordinator facilitates protocol orchestration under the semi-honest adversary model.

Tools and Technologies:

- **MPC Frameworks**: MP-SPDZ, CECILIA for secure computation protocols.
- **Federated Control**: PySyft for federated orchestration.
- **DP-Optional Layer**: Opacus for hybrid DP-MPC experimentation.
- **Languages and Environment**: Python 3.8 on Ubuntu 20.04.

Security Model: All parties are assumed to be semi-honest. TLS-secured communication is enforced between all entities.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2025
DOI: 10.17148/IJARCCE.2025.14643

V. APPLICATION REQUIREMENTS

To support the proposed system, both hardware and software prerequisites must be fulfilled as detailed below.

A. Hardware Requirements

- **Processor**: Multi-core CPU (Intel i7 or AMD Ryzen 7 equivalent or higher).
- **Memory**: Minimum 16 GB RAM for optimal processing of moderate-sized datasets.
- **Storage**: SSD for fast read/write access during secret sharing and reconstruction.
- **Network**: Secure and stable network infrastructure (preferably private VPN) for data exchange between parties.
- **GPU (Optional)**: CUDA-enabled GPU for acceleration of regression and KNN computations in large dataset.

B. Software Requirements

- **Operating System**: Ubuntu 20.04 LTS or compatible Linux distributions (WSL for Windows).
- **Programming Language**: Python 3.8+.
- **MPC Libraries**: MP-SPDZ, TF Encrypted, CECILIA.
- **Data Libraries**: NumPy, Pandas, Scikit-learn (used for evaluation/comparison).
- **IDE**: Visual Studio Code, Jupyter Notebook for experimentation.
- **Security Protocols**: TLS 1.3 or higher for encrypted communication between parties.
- **Containerization (Optional)**: Docker for scalable and reproducible deployments across institutions.

VI. RESULTS

The results for different MPC-based imputation techniques are shown in Table I below:

Method	MAE	Accuracy	Runtime (s)	Notes
MPC Mean	0.000021	100%	0.29	Fast and scalable
MPC Median	0.000017	100%	1.26	Slightly slower than mean
MPC Regression	0.000021	100%	34.0	Works well for correlated data
MPC KNN	0.000018	100%	1440.0	High accuracy, high cost

VI. CONCLUSION

This survey highlights the viability of Secure Multi-Party Computation (MPC) for imputation in sensitive medical datasets. MPC-based imputation methods, including mean, regression, and median, demonstrate strong accuracy and scalability. While KNN shows promise, its computational cost remains a challenge. Hybrid methods involving MPC and differential privacy, as well as deployment in federated environments, represent promising avenues for future research.
We explored a range of imputation techniques—mean, median, regression, and k-nearest neighbors (KNN)—analyzing their strengths and limitations when implemented within MPC frameworks. Our review revealed that while mean and regression imputation methods offer high efficiency and scalability with minimal deviation from plaintext results, KNN, although accurate, imposes significant computational overhead. The use of MPC ensures that sensitive data remains encrypted or secret-shared throughout the process, significantly reducing the risk of data leakage or inference attacks.
We also examined hybrid approaches that combine MPC with Differential Privacy (DP) and Federated Learning (FL) models. These hybrid techniques offer promising directions for balancing performance and privacy, especially in large-scale, cross-institutional healthcare deployments.

REFERENCES

[1]. J. Jentsch, D. Rotem, and M. Parashar, "Privacy-Preserving Data Imputation via Multi-Party Computation for Medical Applications," *Proc. IEEE Int. Conf. E-health Networking, Application & Services (HealthCom)*, pp. 1–6, 2024.
[2]. S. Das, S. Ghosh, and A. Bhattacharya, "Imputation under Differential Privacy," *arXiv preprint arXiv:2206.15063*, 2022.
[3]. G. Jagannathan and R. Wright, "Privacy-Preserving Imputation of Missing Data," *Data & Knowledge Engineering*, vol. 65, no. 1, pp. 40–56, 2008.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2025

DOI: 10.17148/IJARCCE.2025.14643

[4]. G. Gürsoy, M. Brudno, and E. Z. Gursoy, "Privacy-Preserving Genotype Imputation in a Federated Environment Using Fully Homomorphic Encryption," *Cell Systems*, vol. 13, no. 3, pp. 172–186.e6, 2022.

[5]. S.G. Liao, J. Li, and R. S. Brady, "Missing Value Imputation in High-Dimensional Biomedical Datasets: A Comparative Study," *BMC Bioinformatics*, vol. 15, no. 1, pp. 1–16, 2014.

[6]. C. Dwork, A. Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, nos. 3–4, pp. 211–407, 2014.

[7]. N. Dowlin, R. Gilad-Bachrach, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing, "Cryptonets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy," *Proc. Int. Conf. Machine Learning (ICML)*, pp. 201–210, 2016.

[8]. M. Mohassel and Y. Zhang, "SecureML: A System for Scalable Privacy-Preserving Machine Learning," *IEEE Symposium on Security and Privacy*, pp. 19–38, 2017.

[9]. Narayanan, V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," *Proc. IEEE Symp. on Security and Privacy*, pp. 111–125, 2008.

[10]. R. Shokri and V. Shmatikov, "Privacy-Preserving Deep Learning," *Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS)*, pp. 1310–1321, 2015.