International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

Volume 10, Issue 7, May-June-2024 | Published Online : 20th June 2024

Enhanced Privacy in Collaborative Data Collection with Sparse Healthcare Data

M.Blessa Binolin Pepsi ME, S.M.Abinaya, M.Rajshre, M.Siva Darshini

Department of Information Technology, Mepco Schlenk Engineering College, India

ABSTRACT

Privacy-preserving data mining techniques play a crucial role in analyzing diverse information, such as Internet of Things data and healthcare data. However, gathering a substantial amount of sensitive personal information poses a challenge. Additionally, this detail may have missing values, which current methods for collecting personal information do not addresswhile ensuring data privacy. Failing to consider missing values decreases the accuracy of data analysis. In this article, we propose a method for privacy-preserving data collection that accounts for many missing values. The patient data is anonymized and sent to a data collection server. The client then uses the Harris hawk optimization algorithm integrated with particle swarm optimization to determine which indices should be taken and passed on to the server side, reducing computation power and increasing accuracy. The server for data collection constructs a generative model and contingency table specifically designed for multi-attribute analysis, employing expectation-maximization and Gaussian copula methodologies. An efficient server and client architecture is implemented to increase the performance and security of the system. Using differential privacy as a privacy metric, we conduct experiments on synthetic healthcare data, including COVID-19-related data. The results show a 80 – 90% accuracy compared to existing methods that do not consider missing values.

Keywords : Gaussian Copula , Expectation maximization , Differential Privacy , Harris Hawk

I. INTRODUCTION

To effectively manage a pandemic like COVID-19, crucial information such as age, gender, family structure, and medical history of infected individuals is required [1], [2]. While patients may provide such data to medical institutions, it is highly sensitive. Anonymizing this information would allow it to be shared among researchers globally without revealing patient identities, aiding in understanding the pandemic's status and predicting its trajectory more accurately. Even after anonymization, acquiring a substantial amount of sensitive personal information can be challenging. Additionally, this data often contains missing values, as some individuals are willing to provide only partial information. Various methods have been proposed to collect personal information while ensuring data privacy, with many of them relying on differential privacy as the privacy model [3], [4], [5], [6]. However, these methods often overlook missing values, leading to a significant reduction in

data analysis accuracy, especially in multi-attribute analysis scenarios [7], [8].

In this paper, we introduce a method for privacy-preserving data collection that addresses the challenge of missing values. Medical information is rendered anonymous within the patient's personal device or computer within authorized healthcare facilities prior to transmission to a centralized data repository. Each individual retains the autonomy to select the specific data elements they wish to disclose, ensuring greater control over their privacy. Subsequently, the data collection server employs expectation- maximization and Gaussian copula techniques to generate a model and contingency table tailored for multi-attribute analysis. Our approach recognizes that restoring the value distribution of one or two attributes can limit errors in each attribute, even when multiple values are missing. By leveraging copula, which enables data generation based on available information like correlation and mutual information, we combine copula features with data recovery using differential privacy. This innovative approach improves the confidentiality of data collection by employing a copula model in differentially private data collection, particularly in scenarios with numerous missing values. This represents a substantial technical advancement.

Patients furnish their data, encompassing age, gender, and medical history, to authorized entities such as hospitals. Patients have control over what information is shared with the data collection server, which is then made available to researchers worldwide. Patients may also volunteer supplementary information, such as family composition and income, which is not divulged to hospitals. The inclusion of detailed patient data could potentially result in patient identification by researchers even after the removal of all identifiers. However, insufficiently detailed information significantly reduces the effectiveness of data analysis. To tackle this challenge, we suggest the implementation of a differential privacy model. Hospitals utilize a mechanism of differential privacy to handle patient-provided information, encompassing any supplementary data. The information, processed differentially private, along with the additional differentially private data, is then transmitted to the data collection server, which consolidates this data from various hospitals. The server then constructs a generative model to

approximate a model created using true, undisclosed information. Researchers can utilize this generative model to create a contingency table, mine association rules, and employ machine learning techniques, such as deep neural networks, for analysis.

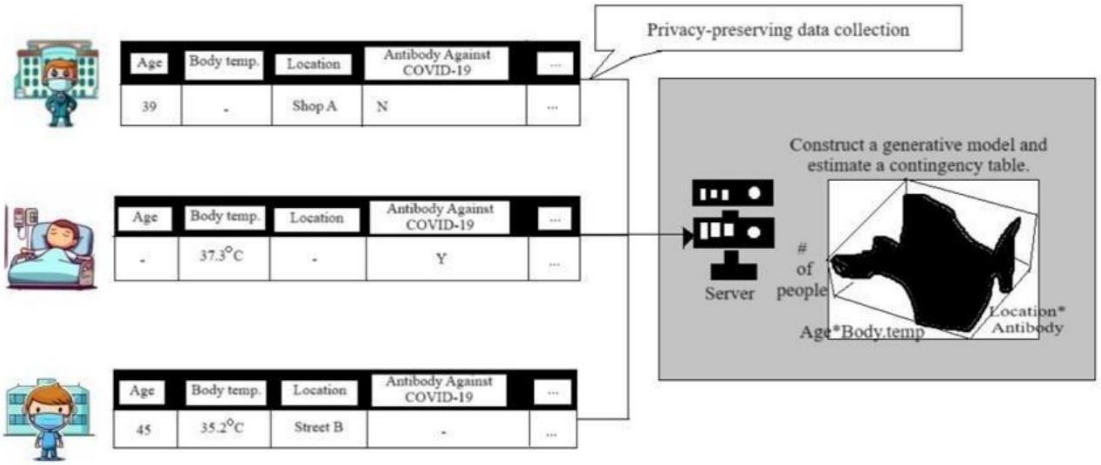


Fig. 1

The Fig.1 portrays a process for collecting data related to COVID-19 while preserving privacy. It features individuals, each associated with their personal details as data points. Data is anonymized as it moves toward a central point called server, and a contingency model is constructed based on this anonymized data.

Moreover, it's essential to consider the scalability and efficiency of the proposed method. Large-scale data collection and analysis require robust systems capable of handling vast amounts of information efficiently. Therefore, our approach incorporates scalable algorithms and distributed computing techniques to ensure timely processing of data from various sources. Additionally, continuous evaluation and refinement of the privacy-preserving mechanisms are crucial to adapt to evolving threats and regulatory requirements in the healthcare data landscape. Thus, our method emphasizes iterative improvement and collaboration with stakeholders to maintain the balance between data utility and privacy protection.

Furthermore, it's imperative to acknowledge the ethical considerations surrounding data collection and privacy preservation in healthcare settings. While the aim is to leverage patient data for public health research and policy-making, it's crucial to prioritize the protection of individuals' privacy rights and maintain trust in the healthcare system. Transparency and informed consent play pivotal roles in ensuring that patients understand how their data will be used and have control over its dissemination. Additionally, efforts should be made to minimize the risk of re-identification, such as implementing robust anonymization techniques and regularly auditing data handling processes. Moreover, fostering collaboration between researchers, policymakers, and advocacy groups can help establish guidelines and best practices for ethical data sharing and governance. By promoting transparency, accountability, and respect for individuals' autonomy, we can uphold the ethical principles underpinning data-driven healthcare initiatives and foster public confidence in the responsible use of sensitive health information.

The paper is arranged as follows: Background, Related works, Proposed Method, Evaluation, Discussion, Conclusion.

2 BACKGROUND

In the context of COVID-19, we assume that all patients share their information with the data collection server through authorized entities like hospitals, mirroring the approach taken in COVID-19 contact tracing applications [9]. For instance, the Ministry of Health, Labor, and Welfare in Japan introduced the COVID-19 Contact-Confirming Application (COCOA), where a user who tests positive for COVID-19 receives a code from an authorized health center to input into COCOA, allowing only users with valid codes to register their infection.

Our proposed approach can be expanded to other contexts, such as crowd-sensing applications, where participants contribute data such as location and accelerometer readings. Since smartphones are capable of health monitoring and cognitive function assessment [10], they can serve as a portal for medical information to the data collection server. In cases where involvement of authorized entities is challenging, incentive and trustworthiness mechanisms from prior studies [11], [12] can be applied.

We also consider the presence of many missing values in the collected data, reflecting the hesitation of many individuals to provide complete information [13], [14]. The rate of missing values is reported to range between 25% and 55%, and may be even higher [15]. Additionally, we assume that the data collection server is honest-but-curious, meaning it follows the proposed scheme honestly but aims to reveal as much personal data as possible. The server is also assumed to

construct a generative model and a contingency table, requiring categorical attribute values. Numerical values are classified into predefined categories in advance.

While some studies on privacy protection assume that users seek services from a service provider based on their attribute values, in our case, individuals voluntarily supply anonymized values to the data collection server and do not anticipate services based on these values, though the server may provide incentives such as financial rewards. The primary objective of the data collection server is to construct a dataset that can be statistically analyzed without requiring precise information about each individual's attribute values. The method incorporates the Harris hawk optimization algorithm for selecting indices and the implementation of an efficient server and client architecture for improved performance and security.

Incorporating the Harris hawk optimization algorithm, our method optimizes the selection of indices to be passed to the server, reducing computation power and enhancing accuracy. This approach is especially beneficial in scenarios with large datasets or complex computations. Furthermore, we implement an efficient server and client architecture to bolster the system's performance and security. This architecture ensures smooth data transmission between the client and server, enhancing the overall user experience and protecting sensitive information.

3 RELATED WORK

3.1 Differential Privacy

Extensive research in the field of data mining [16], [17] has focused on differential privacy models [7].

$$P(A(s_1) \in R) \leq e^{\epsilon} * P(A(s_2) \in R). \quad (1)$$

The equation (1) encapsulates the concept of ε-differential privacy, a cornerstone of privacy-preserving data analysis where A represents a randomized algorithm, s_1 and s_2 denote individual values or data points in the dataset, R represents potential outcomes of the algorithm A, $P(A(s_1) \in R)$ denotes the probability that the outcome R falls within some specified range R when the algorithm A is applied to data point s_1, $P(A(s_2) \in R)$ denotes the probability that the outcome R falls within the same specified range R when the algorithm A is applied to data point s_2 and ε represents a parameter that quantifies the level of privacy protection provided by the algorithm A. It is a measure of how much the output of the algorithm can reveal about any individual data point.

The inequality expresses that for any pair of data points s_1 and s_2, and for any possible outcome R of the algorithm A, the probability that the outcome R falls within the specified range R when applied to data point s_1 is at most e^{ϵ} times the probability of the same outcome R when applied to data point s_2.

In simpler terms, ε-differential privacy ensures that the likelihood of observing a particular outcome from the algorithm A for any individual data point is not significantly affected by the presence or absence of any other individual data point. This property enables individuals to share their data for analysis without fear of their privacy being compromised, as the risk of identifying any single individual's contribution to the data is controlled within the bounds set by ε.

3.2 Anonymized Data Analysis With Differential Privacy

One application of transformed data collection is mobile crowd-sensing, where privacy-preserving techniques can encourage participation [18]. Erlingsson et al. [19] introduced a privacy-preserving method called Randomized Aggregatable Privacy-Preserving Ordinal Response (RAPPOR). Kairouz et al. [3] further analyzed and extended this method, creating O-RAPPOR, which outperformed previous algorithms. O-RAPPOR uses Bloom filters [20], which are modified randomly to ensure differential privacy. However, these methods do not address missing values, requiring the removal of records with missing values beforehand.

Sei et al. [4] introduced S2Mb, an enhancement of the randomized response scheme [21], which offers a technique for estimating accurate counts from values affected by multiple errors [5]. Their focus was on a single attribute without missing values, and they transformed multiple attributes without missing values into a single attribute beforehand. Numerous other studies have also addressed privacy-preserving data collection [22], [23], [24], [25], [26]. Nevertheless, differentially private anonymized data collection methods are greatly impacted by the number of records in the database, leading to a notable decrease in accuracy, especially with small record counts.

Wang et al. introduced a platform for differentially private deep neural networks tailored for handling sensitive crowd-sourced data [27].This platform trains a deep neural network model using sensitive data and publishes the trained model. However, attackers can use model inversion [28], [29] and membership inference attacks [30], [31] to infer sensitive raw data from the trained model. To protect against these attacks, the platform adds noise during the training phase. Nevertheless, Wang et al. assumed that the platform is a trusted entity capable of collecting true information about sensitive data.

3.3 Missing Value Imputation

Wei et al. conducted a comparative analysis assessing the imputation accuracies of eight different methods [32], determining that among them, random forest and quantile regression imputation demonstrated superior performance, particularly for handling left-censored data. In another investigation, Deb and Liew introduced a technique tailored for handling traffic accident data [33]. Their method employs a decision tree to identify a cluster of associated records, from

which missing values are estimated based on the correlation between absent and present attributes. This approach also encompasses the process of sampling numerous potential estimated values characterized by significant similarity.

Many estimation techniques utilize fuzzy clustering algorithms. Rahman et al. proposed an estimation framework based on fuzzy expectation-maximization and fuzzy clustering for missing values [34]. This method identifies and utilizes records with the highest resemblance to the record with missing values, employing a general fuzzy c-means clustering algorithm for the search. The missing values are then estimated using a fuzzy expectation-maximization algorithm, a variation of the standard expectation-maximization algorithm. Similarly, Sefidian and Daneshpour suggested the Gray-based fuzzy c-means and mutual information feature selection estimation method [36]. This approach calculates the distance between records using the gray relational grade during the clustering algorithm, selecting highly related attributes based on mutual information. Raja and Thangavel presented a rough k-means centroid-based estimation method [37] that can handle inconsistencies and uncertainties in datasets, demonstrating its superior performance compared to simple k-means and fuzzy c-means clustering methods.

All of the previously mentioned methods assume that the acquired values accurately represent the true values. However, this study posits that the server receives disguised data due to differential privacy techniques altering the true values. The nature of these concealed values is contingent upon the differential privacy techniques employed, such as a Bloom filter or a set of dummy values. Therefore, current missing value estimation methods are not suitable in differential privacy scenarios.

4 PROPOSED METHOD

In our method, we utilize differential privacy to anonymize personal data of patients stored on the client's end before transmitting it to the server. The server then gathers the anonymized data and reconstructs the distributions of each attribute, along with all combinations of two attributes. From these distributions, the mutual information of all pairs of attributes is computed. Subsequently, a generative model of the patient personal data is derived from this mutual information using a Gaussian copula [38], [39]. This approach, which relies solely on information about attribute pairs, is resilient to missing values. To visually represent the generative model, a contingency table is created from the generative model and the distribution of each attribute.

For the analysis of the collected differentially private data, the server develops a copula model that mitigates the noise introduced by the differentially private technique. Constructing a copula model necessitates the value distribution of each attribute and mutual information about all attributes. Therefore, our method initially estimates single-attribute distributions and subsequently estimates attribute-pair distributions. This copula model can generate an arbitrary number of data samples without missing values are used to construct a contingency table. Additionally, the Harris hawk optimization algorithm is employed to enhance the selection of attribute indices and reduce computation, further improving the accuracy of our method We utilize the Harris hawk optimization algorithm [45], [46] to enhance the selection of attribute indices, thereby reducing computation and improving accuracy. This algorithm optimizes the selection process, determining which indices should be passed to the server for further analysis. On the server side, the collected data is processed to reconstruct attribute distributions and calculate mutual information for all pairs of attributes. This information is then utilized to construct a generative model using a Gaussian copula. This model has the capability to generate synthetic data samples that closely mirror the original data distribution, even in the presence of missing values.

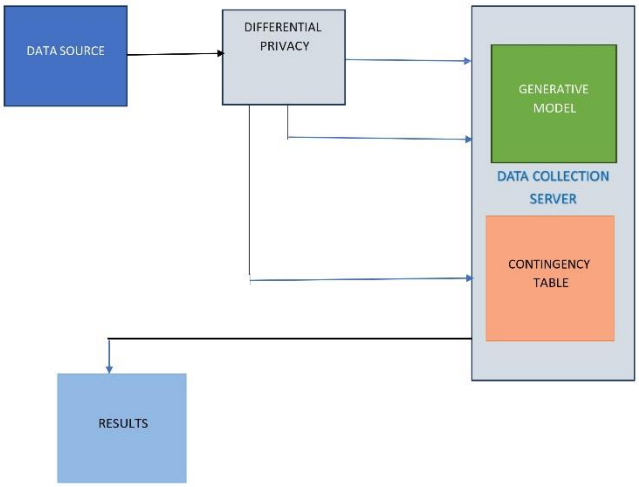


Fig. 2

The Fig.2 depicts data privacy process. It begins with a collected data from the patients, which undergoes anonymization. From there, it leads to differential privacy. The anonymized data is then analyzed using Expectation-Maximization and Gaussian Copula Methods. These processes contribute to the creation of a generative model, which is part of a larger system called the data collection server which also consists of the contingency table.

To handle missing values, our method focuses on the combination of attribute pairs, ensuring robustness in data analysis. The copula model generated by our approach can efficiently produce data samples without missing values, allowing for the construction of a comprehensive contingency table for visualization and further analysis.

Overall, our method offers a comprehensive approach to anonymizing and analyzing patient data, ensuring privacy while maintaining the integrity and accuracy of the data for research and analysis purposes.

4.1 Anonymization at the Client Side

$$R_{ij} = \begin{cases} \{s_{ij}\} \cup Ran(V_j \setminus \{s_{ij}\}, h_j - 1) \text{ with prob. } p_j \\ Ran(V_j \setminus \{s_{ij}\}, h_j) \text{otherwise,} \end{cases} \quad (2)$$

The equation (2) outlines the process of creating a value set Rij for each attribute Aj during the anonymization of patient attribute values where sij denotes the value of attribute Aj for patient i, fj represents the number of categories for attribute Aj, Vj signifies the domain of attribute Aj and Vjk represents the kth value in the domain Vj.
It consists of two cases:

If the attribute value sij is present (i.e., not missing), it remains unchanged. Additional values are selected randomly from the domain Vj (excluding sij). The number of additional values selected is $hj-1$.

If the attribute value sij is missing, a random value from the domain Vj (excluding sij) is selected with probability p_j.

The function $Ran(S,h)$ randomly selects h elements without duplication from the set S. For example, if $S=\{A,B,C\}$ and $h=2$, $Ran(S,h)$ could output sets like $\{A,B\}$, $\{B,C\}$, or $\{A,C\}$.

This anonymization process ensures that each non-missing attribute value is replaced or supplemented with appropriate values from the domain, maintaining the integrity of the data while anonymizing it. The parameters hj and pj control the number of additional values selected and the probability of selecting random values, respectively, contributing to the overall anonymization process.

$$h_j = max\left(\left\lfloor \frac{f_j}{1+e^{\epsilon}} \right\rfloor, 1\right) \text{and}$$

$$p_j = \frac{e^{\epsilon} h_j}{f_j - h_j + e^{\epsilon} h_j} \quad (3)$$

The equation (3) outlines the determination of parameters hj and pj to achieve ϵ-differential privacy in the context of anonymizing patient attribute values where hj represents the number of additional values selected when anonymizing attribute Aj, pj denotes the probability of selecting random values when attribute Aj values are missing, fj signifies the number of categories for attribute Aj and ϵ represents the privacy parameter, controlling the level of differential privacy.

The formula for h_j calculates the number of additional values (hj) to be selected for attribute Aj during anonymization. It ensures that the value of hj is at least 1 and takes into account the number of categories (f_j) and the privacy parameter (ϵ).

The formula for p_j calculates the probability (pj) of selecting random values when attribute Aj values are missing.It considers the number of additional values (hj), the number of categories (f_j), and the privacy parameter (ϵ).

Each value set Rij should adhere to ϵ/g-differential privacy, where g represents the number of attributes. The allocation of the privacy budget (ϵ/g) ensures that privacy guarantees are evenly distributed among attributes.Even when attributes are identical, achieving ϵ-differential privacy is possible due to the composition property of differential privacy [40], which allows for the aggregation of privacy guarantees across attributes.

Algorithm 1 runs at client side, which is the anonymization algorithm.

Algorithm 1. Anonymization Algorithm

Input: Privacy budget for the differential privacy $\in$, Original data $s_{i1}, \ldots, s_{ig}$ which is the true attribute value from each jth attribute of A_j from every domain of V_j

Output: Disguised version of Original data, R_i

Step 1: for j = 1 ,..., g (number of attributes) do

Step 2: $f_j \leftarrow |V_j|$,size of V_j is determined

Step 3: Based on (3), determine p_j and h_j by substituting ϵ/g into ϵ

Step 4: Based on (2), obtain R_{ij} from s_{ij} and V_j

Step 5: end for

Step 6: return R_i

4.2 Estimation at the Server Side

The data collection server first estimates the distribution of each attribute's values, as explained in Section 4.2.1. With these estimated distributions, the server builds a generative model, specifically a Gaussian copula (see Section 4.2.2). Next, it produces n complete data records and creates a contingency table of target attributes, as directed by a data analyzer (Sections 4.2.3).

4.2.a) Optimization algorithm :

Integrating the Harris Hawk Optimization (HHO) algorithm with the Particle Swarm Optimization (PSO) technique aims to create a robust optimization strategy that leverages the strengths of both algorithms. The objective is to develop an approach that effectively explores the search space while exploiting promising regions, leading to improved convergence speed and solution quality. Inspired by the hunting behavior of Harris's Hawks, HHO simulates the interaction between prey and predators to update the population's positions iteratively. Meanwhile, PSO draws inspiration from bird flocking behavior, where each particle adjusts its position based on its own experience and the flock's best performer. By combining these algorithms, the objective is to enhance optimization performance, making it applicable to various optimization problems such as function optimization, feature selection, and parameter tuning in machine learning algorithms.

The integration of the Harris Hawk Optimization (HHO) algorithm with the Particle Swarm Optimization (PSO) [47] technique provides a promising optimization strategy. HHO [46] mimics the hunting behavior of Harris's Hawks, updating positions based on prey-predator interactions. In contrast, PSO [48] is inspired by bird flocking, with particles adjusting positions according to personal and collective experiences. By merging these methods, the exploration-exploitation trade-off is balanced, potentially leading to faster convergence and higher-quality solutions. The formulas provided facilitate the implementation of these algorithms, governing particle movements based on their current positions, velocities, and social influences.

```
np.sum(solution ** 2)
```
(4)

The equation (4) calculates the sum of squares of elements in the solution array. It's commonly used in optimization algorithms as part of the fitness evaluation process, where the objective is to minimize this sum to reach the optimal solution.

```
social_weight * r2 * (self.global_best_position - self.particles[i])
```
(5)

The equation (5) represents the update rule for the velocity component in the PSO algorithm. It adjusts the velocity of each particle based on the difference between the global best position (self.global_best_position) and the current position of the particle (self.particles[i]), scaled by the social weight and a random factor r2.

```
move_vector = self.hawks[rand_hawk_idx] - self.hawks[i]
self.hawks[i] += move_vector * np.random.rand()
```
(6)

The equation (6) calculates the movement vector for each hawk (move_vector) based on the difference between the position of a randomly selected hawk (self.hawks[rand_hawk_idx]) and the current hawk's position (self.hawks[i]). Then, the position of the current hawk is updated by adding a fraction of the movement vector, scaled by a random factor. This process mimics the hunting behavior of hawks in the HHO algorithm.

4.2.1 Separated Estimation: Estimation of a Value Distribution of Each Attribute :

$$q_i = \frac{p_j(h_j-1)}{f_j-1} + \frac{(1-p_j)h_j}{f_j-1} = \frac{h_j-p_j}{f_j-1}, \quad (7)$$

$$\begin{pmatrix} w_{j1} \\ w_{j2} \\ \vdots \\ w_{jf_j} \end{pmatrix} = M \begin{pmatrix} u_{j1} \\ u_{j2} \\ \vdots \\ u_{jf_j} \end{pmatrix}, \quad (8)$$

$$\begin{pmatrix} z_{j1} \\ z_{j2} \\ \vdots \\ z_{jf_j} \end{pmatrix} = M^{-1} \begin{pmatrix} w_{j1} \\ w_{j2} \\ \vdots \\ w_{jf_j} \end{pmatrix} \quad (9)$$

Each client sends its true value and $hj-1$ randomly selected values (other than the true value) with probability p_j for attribute j. The probability of sending another value instead of the true value is calculated using the equation (7) where q_i represents the probability of sending another value instead of the true value.

The equations (8) and (9) depict transformations involving matrices M and M^{-1}. These transformations likely play a role in data processing or analysis.

Due to the limitations of matrix M's estimation accuracy and the computational complexity of its inverse function, an EM-based algorithm is employed. The EM algorithm is utilized for maximum a posteriori estimation, treating certain variables (u_{jk}) as unobserved latent variables. By iteratively estimating the unobserved latent variables (u_{jk}), the EM algorithm identifies the variables that best explain the observed values (w_{jk}).

Expectation-Maximization (EM) algorithm ensures likelihood improvement with each iteration [41], [42]. The equation (10) calculates the estimated count of value occurrences within attributes. The approach combines EM with other techniques for exploration and exploitation. The introduction of z_{jk} denotes estimated value occurrences, enhancing the estimation process.

$$\tilde{n}_j = \sum_{k=1}^{f_j} w_{jk}. \quad (10)$$

$$z_{jk} \leftarrow z_{jk}\,(p_j D_k + q_j(\varepsilon - D_k)), \quad (11)$$

where

$$q_j = \frac{h_j-p_j}{f_j-1}, \quad (12)$$

$$D_k = \frac{w_{jk}}{p_j z_{jk} + q_j(\tilde{n}_j - z_{jk})}, \quad (13)$$

This outlines a process for obtaining estimated occurrences (z_{jk}) of value combinations Vjk in attribute Aj using an expectation-maximization (EM)-based algorithm.

The equation (11) updates the estimated occurrences of Vjk (zjk) iteratively based on substitution. It involves a weighted sum of two terms, with coefficients $p_j D_k$ and $q_j(\varepsilon - D_k)$, influencing the update.

The equation (12) defines qj, representing the probability of selecting a value other than the true one from the domain Vj. It depends on parameters h_j, p_j, and f_j, reflecting the number of categories, the probability of selecting the true

value, and the total number of categories for attribute Aj, respectively.

The equation (13) computes D_k, which involves the ratio of observed occurrence (w_{jk}) to the expected occurrences under the current estimate of z_{jk}. It incorporates parameters h_j, p_j, q_j and $\tilde{n}$, where $\tilde{n}$ represents the estimated count of occurrences for attribute A_j.

This refers to estimating the occurrence of each combination of attribute values for n patients, enabling the estimation of value distributions for all attribute pairs A_j and A_{jf} .These outline a systematic approach for updating estimated occurrences of attribute value combinations using an EM-based algorithm.

Mutual information measures the dependency between two random variables, indicating how much knowing one variable reduces uncertainty about the other. Mutual information quantifies the amount of information obtained about one attribute by observing the other. A higher mutual information value indicates a stronger relationship or dependency between the two attributes. If the mutual information is zero, it suggests that the attributes are independent.

$$\sum_{k\in V_j}\sum_{k^F\in V_{j^F}} p(k,k')\log\frac{p(k,k^F)}{p(k)p(k^F)}, \qquad (14)$$

The equation (14) calculates mutual information by summing over all possible values of attributes Aj and Aj'. For each combination of attribute values Vjk and $Vj'k'$, it computes the joint probability $p(k,k')$ and compares it to the product of the marginal probabilities $p(k)$ and $p(k')$. The logarithm of the ratio of joint probability to the product of marginal probabilities is multiplied by the joint probability, and this product is summed over all possible combinations where $p(k,k')$ represents the joint probability that V_{jk} and $V_{j^Fk^F}$ occur for attribute $A_{j'}$, and $p(k)$ represents the probability that V_{jk} occurs for attribute A_j.

4.2.2 Generative Model Construction: Using a Gaussian Copula Constructing a Generative Model

We created n complete datasets by employing both the Gaussian copula C and the reconstructed data as described in Section 4.2.1. The number of instances n for each attribute Aj was determined based on the estimated attribute distribution outlined in Section 4.2.1. Additionally, random values $x1, \ldots, xg$ were generated from a multivariate Gaussian distribution with a covariance matrix Σ. Subsequently, for each instance $i = 1, \ldots, g$, we computed $ui = \Phi(xj)$. Finally, leveraging the reconstructed data detailed in Section 4.2.1, we obtained the inverse cumulative distribution function F_{uj}^{-1} for each attribute value, where Fj denotes the estimated attribute distribution.

4.2.3 Contingency Table Creation: Counting Each Combination of the Target Attributes

After completing the aforementioned procedure, we obtained n complete data records comprising g attributes. When dealing with a contingency table involving numerous attributes, its core utility may diminish [43], [44]. As a result, analysts typically opt for a subset of attributes. Subsequently, the target contingency table is generated by tallying the occurrences of each combination of attribute values from the n complete data records generated.

To mitigate computational demands and enhance precision, the client employs an amalgamated version of Particle Swarm Optimization and Harris Hawk Optimization algorithms [45] to select and relay indices to the server. The data collection server formulates a generative model and contingency table for multi-attribute analysis utilizing expectation-maximization and Gaussian copula techniques.

A streamlined server-client architecture is implemented to bolster system efficiency and security. Leveraging differential privacy as a privacy metric, experiments are conducted on synthetic healthcare data. This approach is instrumental in fortifying the model's resilience, thereby facilitating the generation of an effective contingency model.

Algorithm 2 runs at server side.

Algorithm 2. Creation of Gaussian Coupla and Contingency Table

Input: Privacy Parameter $\in$, From (3) the parameters p_j and h_j , algorithm 1 result R_i and a set of target names for contingency table

Output: Contingency table for the target attributes

Step 1: for j = 1,…,g (number of attributes) do

Step 2 : $Z_j \leftarrow$ Based on Equation (8) , which is estimated value distribution of A_j

Step 3: end for

Step 4: for j = 1,…,g do

Step 5: for j' = 1,…,g do

Step 6: if j $\neq$ j' then

Step 7: $Z_{jj'} \leftarrow$ Based on Equations (12) , (13), and (8) which is estimated value distribution of combination of A_j and $A_{j'}$

Step 8: end if

Step 9: end for

Step 10: end for

Step 11: for j = 1,…,g do

Step 12: Construct the cumulative distribution function F_j

Step 13: end for

Step 14: for j = 1,…,g do

Step 15: for $j' = 1,\ldots,g$ do

Step 16: if j ≠ j' then

Step 17: do change of $\sum$ converges

Step 18: Cumulative distribution function F_j , $F_{j'}$ and $\sum$ (maximum likelihood estimator) From Equation (10)

Step 19: Based on the mutual information of $Z_{j\,j'}$

Step 20: end for

Step 21: end if

Step 22: end for

Step 23: end while

Step 24: $O \leftarrow \emptyset$ (standard Gaussian distribution)

Step 25: for i = 1,…,n do

Step 26: $\{ x_1, .x_j \} \leftarrow$ gaussian distribution with covariance matrix

Step 27: for j = 1,…,g do

Step 28: $u_j \leftarrow \Phi(x_j)$ represents the probability density function of a standard Gaussian distribution

Step 29: $z_j \leftarrow F_j^{-1}(u_j)$ represents the marginal distribution of attribute A_j

Step 30: $O \leftarrow O \cup \{\{z_1, \ldots, z_g\}\}$

Step 31: end for

Step 32: end for

Step 33: return Contingency table for the target attributes with the combination of attribute values in O.

5 EVALUATION

5.1 Evaluation Setting

In this study, a contingency table is considered as a representation of the probability distribution of attribute values. To measure the difference between these distributions, the Jensen–Shannon (JS) divergence was chosen over the more commonly used Kullback–Leibler (KL) divergence. This decision was made because the KL divergence requires all probabilities to be non-zero; if any probabilities are zero, the KL divergence cannot be calculated due to division-by-zero errors. In contrast, the JS divergence, which is derived from the KL divergence, does not have this constraint.

The study tested various missing value rates, m, ranging from 0.3 to 0.8, and considered analyses with different numbers of attributes, c, ranging from 1 to 5. The reported results are averages of 100 experiments for each setting, with default parameters m=0.5, c=3, and ϵ=5. It is important to note that the missing value rate m is used only for experimental purposes, and the proposed algorithm is not dependent on it. The number of attributes, c, targeted for analysis can be chosen freely by the data analyst based on the specific goals of the analysis. In our healthcare project, you are addressing the challenge of missing values in synthetic diabetes, COVID-19, and heart disease prediction datasets, while creating a generative model. To handle missing values, you can employ advanced machine learning techniques such as Generative Adversarial Imputation Nets (GAIN), which has shown promising results in imputing missing data accurately and efficiently. GAIN has been proven to outperform other commonly used methods like MICE and missForest, particularly when dealing with high missingness rates and skewed or imbalanced variables.

In addition to GAIN, you can also explore the use of other methods tailored for mixed-type datasets, such as missForest, which is based on the random forest algorithm. Although missForest has demonstrated superior performance in certain scenarios, it may suffer from long computation times, making it less practical for big data research.

By employing these advanced techniques, you can ensure that your generative model is trained on high-quality, complete datasets, ultimately improving its predictive capabilities and generalizability in healthcare applications. This approach not only addresses the pervasive problem of missing data but also leverages the power of machine learning to create more robust and reliable models for diabetes, COVID-19, and heart disease prediction.

In this specific synthetic diabetes, COVID-19, and heart disease datasets, JS divergence can be employed to assess the similarity between the original and synthetic datasets. By comparing the distributions of various features, such as age, gender, or clinical measurements, researchers can ensure that the synthetic datasets maintain the essential characteristics of the original datasets.

5.2 Evaluation Metrics

To thoroughly assess the algorithm's efficiency, we utilized the following performance evaluation metrics:

Accuracy: Accuracy measures the proportion of correctly classified instances among all instances in the dataset. It's a general measure of the model's correctness and is calculated as:

$$\text{Accuracy} = \frac{True\ positives + True\ Negatives}{Total\ Population}$$

5.3 Results

TABLE 1
Informative Data

Healthcare dataset	Records	Features	Noise Percentage
Heart	299	13	12.80%
Stroke	584	12	26.60%
Diabetes	768	9	14.77%
COVID-19	569	32	19.36%
Breast Cancer	188	11	18.85%

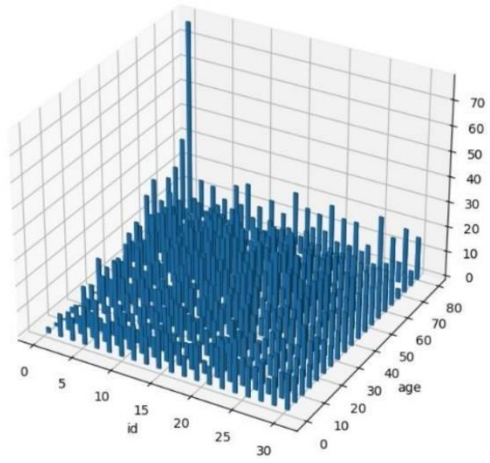
Heart: It includes various attributes like age, anaemia, creatinine_phosphokinase,diabetes,ejection_fraction, high_blood_pressure,platelets,serum_creatinine, serum_sodium, sex, smoking, time and the target attribute death_event.

Stroke: It includes attributes like id, gender, age, hypertension, heart_disease, ever_married, work_type, residence_type, avg_glucose_level, bmi, smoking_status and target attribute as stroke.

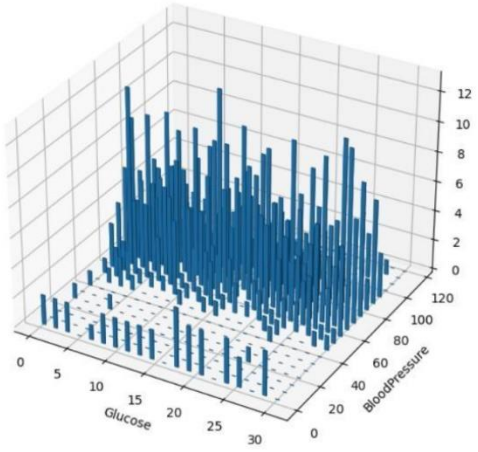
Diabetes: It includes attributes like Pregnancies, Glucose, Blood Pressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction,Age and target attribute as Outcome.

COVID-19: It include attributes like Confirmed, Deaths, Recovered, Active, New cases, New deaths, New recovered, Deaths/100 cases, Recovered/100 cases, Deaths/100 recovered with target attribute as no. of countries.

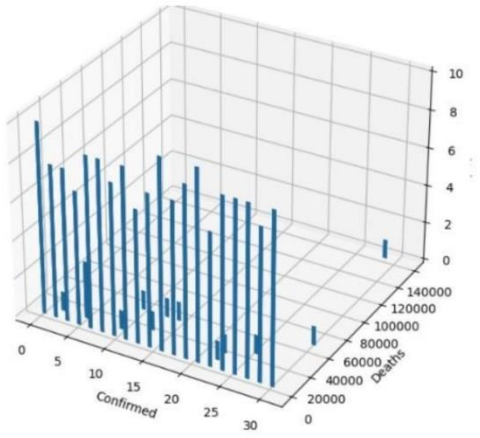
Breast Cancer: It includes attributes like radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean and target attribute as diagnosis.



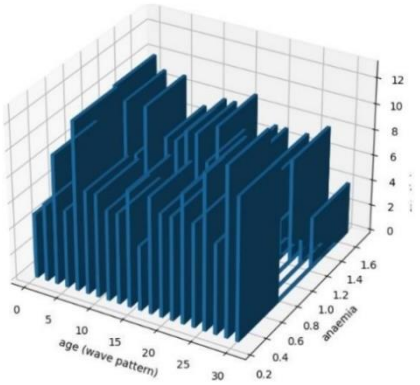
(a)



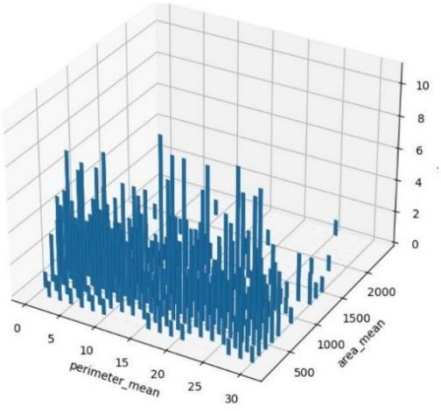
(b)



(c)



(d)



(e)

a) **Stroke**: This graph depicts how the dataset stroke has changed with the addition of white noise after the process of anonymization with id and age being plotted against frequency.
b) **Diabetes**: This graph depicts how the dataset diabetes has changed with the addition of white noise after the process of anonymization with Glucose and BloodPressure being plotted against frequency.
c) **COVID-19**: This graph depicts how the dataset COVID-19 has changed with the addition of white noise after the process of anonymization with a plot showing confirmed against deaths with frequency.
d) **Heart**: This graph depicts how the dataset heart has changed with the addition of white noise after the process of anonymization age and anaemia being plotted aginst frequency.
e) **Breast Cancer**: This graph depicts how the dataset breast cancer has changed with the addition of white noise after the process of anonymization with perimeter_mean and area_mean plotted aginst accuracy.

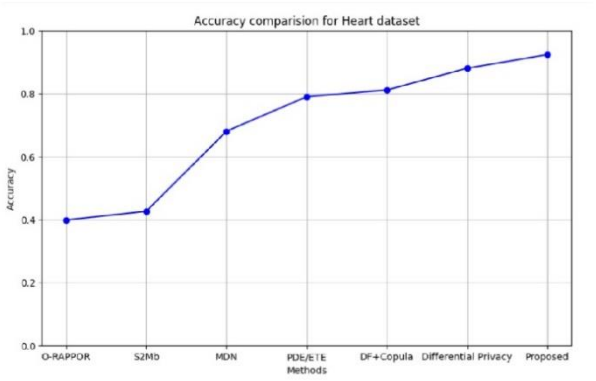


Fig. 3

Fig. 3 depicts the comparison of accuracy with our proposed method to all the other methods for Heart dataset showing higher accuracy for our result.

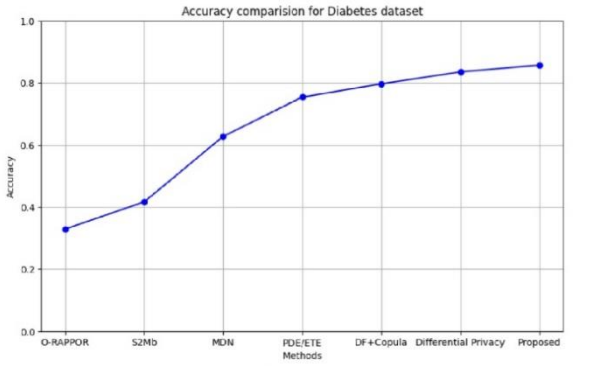


Fig. 4

Fig. 4 depicts the comparison of accuracy with our proposed method to all the other methods for Diabetes dataset showing higher accuracy for our result.

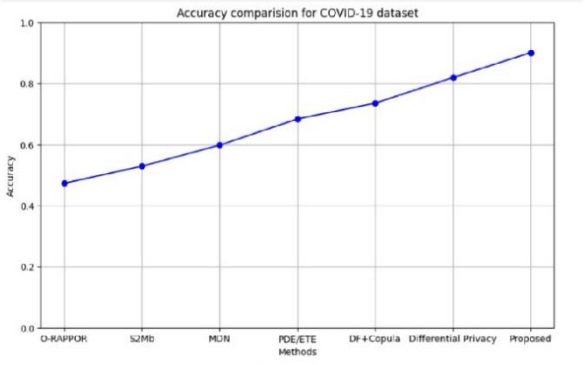


Fig. 5

Fig. 5 depicts the comparison of accuracy with our proposed method to all the other methods for COVID-19 dataset showing higher accuracy for our result.

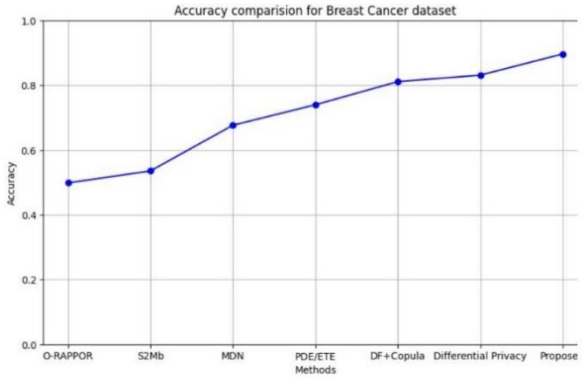


Fig. 6

Fig. 6 depicts the comparison of accuracy with our proposed method to all the other methods for Breast Cancer dataset showing higher accuracy for our result.

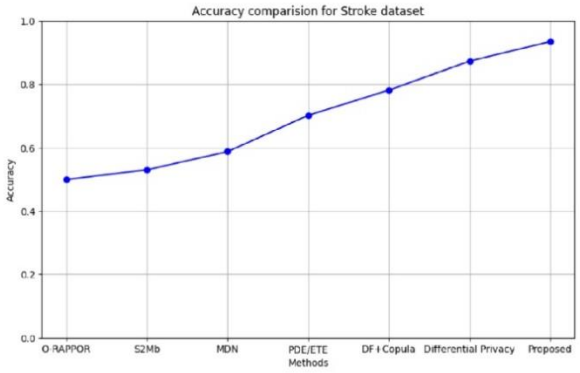


Fig. 7

Fig. 7 depicts the comparison of accuracy with our proposed method to all the other methods for Stroke dataset showing higher accuracy for our result.

The other methods include O-RAPPOR, S2Mb, MDN, PDE/ETE, DF+Copula, Differential Privacy. Comparing with these methods, our proposed method produces higher accuracy for all healthcare datasets.

6 DISCUSSION

6.1 Extensions of ∈-Differential Privacy

This study primarily concentrates on ϵ-differential privacy and ϵ-local differential privacy. However, several modifications and expansions of ϵ-differential privacy (and ϵ-local differential privacy) have been proposed in the literature. These extensions include Gaussian differential privacy, concentrated differential privacy, Bayesian differential privacy, and Renyi differential privacy. Despite these extensions, many studies, especially those involving differentially private federated learning (which constructs a machine-learning model based on distributed data), still predominantly aim for ϵ- or (ϵ, δ)-differential privacy.

ϵ-Differential privacy is considered the foundational concept underpinning various definitions of differential privacy, offering stronger privacy assurances compared to these relaxation techniques. Consequently, ϵ-differential privacy remains a popular choice in recent research. Hence, our focus in this study remains on ϵ-differential privacy (and ϵ-local differential privacy). In future investigations, we may delve into (ϵ, δ)-differential privacy and other extensions.

7 Conclusion

Patient information, crucial for monitoring infections like COVID-19, is often shared with researchers. While privacy protection is vital, an excessive focus on it can hinder data analysis. Many patients withhold personal information or provide only some attributes due to privacy concerns, resulting in datasets with numerous missing values. Existing privacy-protection data mining methods often overlook these missing values, significantly reducing data analysis accuracy.

This study presents a novel approach that deduces the value distributions of individual attributes and pairs of attributes to construct a Gaussian copula. By leveraging information from attribute combinations, this method proves robust against missing values. The constructed Gaussian copula integrates information from all pairs of attributes, thereby improving data reproducibility. Through the utilization of actual COVID-19 data, we illustrate that our approach substantially diminishes the Jensen-Shannon divergence in comparison to existing methods.

While this study assesses the proposed method using publicly available data, future endeavors aim to gather more sensitive attribute values utilizing this approach. Additionally, this study integrates the Particle Swarm Optimization (PSO) and Harris Hawk (HH) algorithms on the client-side to select essential attributes, reducing computation power and increasing accuracy in data analysis.

In future endeavors, the proposed method could be extended to incorporate real-world healthcare datasets beyond COVID-19 data, potentially through collaborations with healthcare institutions to gather more comprehensive and diverse patient attribute values while ensuring privacy protection. Additionally, efforts could focus on adapting the method to handle more complex data structures, such as longitudinal or hierarchical data, to increase its applicability across various healthcare domains. Optimizing the computational efficiency of the Particle Swarm Optimization (PSO) and Harris Hawk (HH) algorithms, or exploring alternative optimization techniques, would enhance scalability for larger datasets. Evaluating the method across diverse healthcare domains beyond infectious diseases, validating its effectiveness using independent datasets, and assessing its generalizability across different populations and geographic regions would further validate its utility. Integration with additional privacy-preserving techniques and the development

of user-friendly software tools or platforms would facilitate widespread adoption and ensure robust data security and privacy protection in real-world healthcare applications.

REFERENCES

[1]C. Mazza et al., "A nationwide survey of psychological distress among italian people during the COVID-19 pandemic: Immediate psychological responses and associated factors," Int. J. Environ. Res. Public Health, vol. 17, no. 9, 2020, Art. no. 3165.

[2] N. W. Chew et al., "A multinational, multicentre study on the psychological outcomes and associated physical symptoms amongst healthcare workers during COVID-19 outbreak," Brain, Behavior, Immun., vol. 88, pp. 559–565, 2020.

[3] P. Kairouz, K. Bonawitz, and D. Ramage, "Discrete distribution estimation under local privacy," in Proc. 33rd Int. Conf. Mach. Learn., 2016, pp. 2436–2444.

[4] Y. Sei and A. Ohsuga, "Differential private data collection and analysis based on randomized multiple dummies for untrusted mobile crowdsensing," IEEE Trans. Inf. Forensics Secur., vol. 12, no. 4, pp. 926–939, Apr. 2017.

[5] Y. Sei and A. Ohsuga, "Differentially private mobile crowd sensing considering sensing errors," Sensors, vol. 20, no. 10, pp. 2785:1–2785:25, 2020.

[6] J. Xu, A. Wang, J. Wu, C. Wang, R. Wang, and F. Zhou, "SPCSS: Social network based privacy-preserving criminal suspects sensing," IEEE Trans. Computat. Social Syst., vol. 7, no. 1, pp. 261–274, Feb. 2020.

[7] C. Dwork, "Differential privacy," in Proc. Int. Colloq. Automata Lang. Program., 2006, pp. 1–12.

[8] A. Roy Chowdhury, C. Wang, X. He, A. MacHanavajjhala, and S. Jha, "Crypt: Crypto-assisted differential privacy on untrusted servers," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2020, pp. 603–619.

[9] J. Abeler, M. B€acker, U. Buermeyer, and H. Zillessen, "COVID-19 contact tracing and data protection can go together," JMIR mHealth uHealth, vol. 8, no. 4, 2020, Art. no. e19359.

[10] E. L. Brown, N. Ruggiano, J. Li, P. J. Clarke, E. S. Kay, and V. Hristidis, "Smartphone-based health technologies for dementia care: Opportunities, challenges, and current practices," J. Appl. Gerontol., vol. 38, no. 1, pp. 73–91, 2019. [Online]. Available: https:// pubmed.ncbi.nlm.nih.gov/28774215/

[11] M. Pouryazdan, B. Kantarci, T. Soyata, L. Foschini, and H. Song, "Quantifying user reputation scores, data trustworthiness, and user incentives in mobile crowd-sensing," IEEE Access, vol. 5, pp. 1382–1397, 2017.

[12] A. Suliman, H. Otrok, R. Mizouni, S. Singh, and A. Ouali, "A greedy-proof incentive-compatible mechanism for group recruitment in mobile crowd sensing," Future Gener. Comput. Syst., vol. 101, pp. 1158–1167, 2019.

[13] H. Kurasawa et al., "Missing sensor value estimation method for participatory sensing environment," in Proc. IEEE Int. Conf. Pervasive Comput. Commun., 2014, pp. 103–111.

[14] L. Cheng et al., "Compressive sensing based data quality improvement for crowd-sensing applications," J. Netw. Comput. Appl., vol. 77, pp. 123–134, 2017.

[15] Z. Wu et al., "A location privacy-preserving system based on query range cover-up or location-based services," IEEE Trans. Veh. Technol., vol. 69, no. 5, pp. 5244–5254, May 2020.

[16] M. U. Hassan, M. H. Rehmani, and J. Chen, "Differential privacy techniques for cyber physical systems: A survey," IEEE Commun. Surv. Tuts., vol. 22, no. 1, pp. 746–789, Jan.–Mar. 2020.

[17] P. Zhao, G. Zhang, S. Wan, G. Liu, and T. Umer, "A survey of local differential privacy for securing internet of vehicles," J. Supercomputing, vol. 76, pp. 8391–8412, 2020. [Online].Available:https://link.springer.com/article/10.1007/s11227–019-03104-0

[18] Y. Wang, Z. Cai, Z. H. Zhan, B. Zhao, X. Tong, and L. Qi, "Walrasian equilibrium-based multiobjective optimization for task allocation in mobile crowdsourcing," IEEE Trans. Computat. Social Syst., vol. 7, no. 4, pp. 1033–1046, Aug. 2020.

SEI ET AL.: PRIVACY-PRESERVING COLLABORATIVE DATA COLLECTION AND ANALYSIS WITH MANY MISSING VALUES 2171

[19] U. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized aggregatable privacy-preserving ordinal response," in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., 2014, pp. 1054–1067.

[20] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," Commun. ACM, vol. 13, no. 7, pp. 422–426, 1970.

[21] S. Agrawal and J. Haritsa, "A framework for high-accuracy privacy-preserving mining," in Proc. 21st Int. Conf. Data Eng., 2005, pp. 193–204.

[22] M. M. Groat, B. Edwards, J. Horey, W. He, and S. Forrest, "Enhancing privacy in participatory sensing applications with multidimensional data," in Proc. IEEE Int. Conf. Pervasive Comput. Commun., 2012, pp. 144–152.

[23] M. M. Groat, B. Edwards, J. Horey, W. He, and S. Forrest, "Application and analysis of multidimensional negative surveys in participatory sensing applications," Pervasive Mobile Comput., vol. 9, no. 9, pp. 372–391, 2013.

[24] R. Agrawal, R. Srikant, and D. Thomas, "Privacy preserving OLAP," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2005, pp. 251–262.

[25] R. Chaytor and K. Wang, "Small domain randomization: Same privacy, more utility," Proc. VLDB Endow., vol. 3, no. 1/2, pp. 608–618, 2010.

[26] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in Proc. 22nd ACM SIGMODSIGACT-SIGART Symp. Princ. Database Syst., 2003, pp. 211–222.

[27] Y. Wang, M. Gu, J. Ma, and Q. Jin, "DNN-DP: Differential privacy enabled deep neural network learning framework for sensitive crowdsourcing data," IEEE Trans. Computat. Social Syst., vol. 7, no. 1, pp. 215–224, Feb. 2020.

[28] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., 2015, pp. 1322–1333.

[29] 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.
[30] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in Proc. IEEE Secur. Privacy, 2017, pp. 3–18.
[31] G. Liu, C. Wang, K. Peng, H. Huang, Y. Li, and W. Cheng, "SocInf: Membership inference attacks on social media health data with machine learning," IEEE Trans. Computat. Social Syst., vol. 6, no. 5, pp. 907–921, Oct. 2019.
[32] R. Wei et al., "Missing value imputation approach for mass spectrometry-based metabolomics data," Sci. Rep., vol. 8, no. 1,pp.1–10,2018.[Online].Available: https://www.nature.com/ articles/s41598–017- 19120-0
[33] R. Deb and A.W. C. Liew, "Missing value imputation for the analysis of incomplete traffic accident data," Inf. Sci., vol. 339, pp. 274–289, 2016.
[34] M. G. Rahman and M. Z. Islam, "Missing value imputation using a fuzzy clustering-based EM approach," Knowl. Inf. Syst., vol. 46, no. 2, pp. 389–422, 2016. [Online]. Available: https://link.springer.com/ article/10.1007/s10115–015-0822-y
[35] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values," J. Climate, vol. 14, no. 5, pp. 853–871, 2001.
[36] A. M. Sefidian and N. Daneshpour, "Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model," Expert Syst. Appl., vol. 115, pp. 68–94, 2019.
[37] P. S. Raja and K. Thangavel, "Missing value imputation using unsupervised machine learning techniques," Soft Comput., vol. 24, no. 6, pp. 4361–4392, 2020. [Online]. Available: https://link.springer.com/ article/10.1007/s00500–019-04199-6
[38] C. Genest and J. MacKay, "The joy of copulas: Bivariate distributions with uniform marginals," Amer. Statistician, vol. 40, no. 4, pp. 280–283, 1986.
[39] L. Rocher, J. M. Hendrickx, and Y.-A. de Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," Nature Commun., vol. 10, no. 1, pp. 1–9, 2019.
[40] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?," SIAM J. Comput., vol. 40, no. 3, pp. 793–826, 2013.
[41] C. F. J. Wu, "On the convergence properties of the EM algorithm on JSTOR," Ann. Statist., vol. 11, no. 1, pp. 95–103, 1983.[Online].Available:https://www.jstor.org/stable/2240463 ?seq¼1#metadata_info_ tab_contents
[42] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based bias field correction of MR images of the brain," IEEE Trans. Med. Imag., vol. 18, no. 10, pp. 885–896, Oct. 1999.
[43] M. J. Wood and J. Ross-Kerr, Basic Steps in Planning Nursing Research: From Question to Proposal. Burlington, MA, USA: Jones & Bartlett Publishers, 2010.
[44] R. Grover and M. Vriens, The Handbook of Marketing Research: Uses, Misuses, and Future Advances, Thousand Oaks, CA, USA: SAGE Publications, Inc, 2006.
[45] Zenab Mohamad Elgamal, Norizon Binti Mohd Yasin, Mohammad Tubishat, Mohammad Alswaitti, and Seyedali Mirjalili (Senior Member, IEEE), "An Improved Harris Hawks Optimization Algorithm With Simulated Annealing for Feature Selection in the Medical Field," Entropy," IEEE Access, vol. 9, 2020
[46] Pei-Wen Shu, Qing-Xin Chu, and Jian-Ye Mai, "Harris Hawks Optimization Algorithm for Waveguide Filter Designs"in Proc. IEEE Conf. Secur. Privacy, 2020, pp
[47] Zijing Yuan, Jiayi Li, Haichuan Yang, and Baohang Zhang, "A Hybrid Whale Optimization and Particle Swarm Optimization Algorithm"in Proc. IEEE Conf. Secur. Privacy, 2021, pp
[48] Tao Sui, Huimin Cui, Ning Liang, Xiuzhi Liu, Dong Liu, and Qingru Wang, "Research and Algorithm Test of Adaptive Interbreeding Hybrid Particle Swarm Optimization " IEEE Conf. Secur. Privacy, 2020, pp