scientific reports

Check for updates

OPEN

Application of the joint clustering algorithm based on Gaussian kernels and differential privacy in lung cancer identification

Hang Yanping, Zheng Haixia, Yang Minmin, Wang Nan, Kong Miaomiao & Zhao Mingming✉

In the age of big data, privacy, particularly medical data privacy, is becoming increasingly important. Differential privacy (DP) has emerged as a key method for safeguarding privacy during data analysis and publishing. Cancer identification and classification play a vital role in early detection and treatment. This paper introduces a novel algorithm, DPFCM_GK, which combines differential privacy with fuzzy c-means (FCM) clustering using a Gaussian kernel function. The algorithm enhances cancer detection while ensuring data privacy. Three publicly available lung cancer datasets, along with a dataset from our hospital, are used to test and demonstrate the effectiveness of DPFCM_GK. The experimental results show that DPFCM_GK achieves high clustering accuracy and enhanced privacy as the privacy budget (ε) increases. For the UCIML, NLST, and NSCLC datasets, it reaches optimal results at lower ε (1.52, 1.24, and 2.32) compared to DPFCM. In the lung cancer dataset, DPFCM_GK outperforms DPFCM within, $0.05 \le \varepsilon \le 2.5$, with significant differences ($\chi^2 = 4.54 \sim 29.12$; $P < 0.05$), and both methods converge to an accuracy of 94.5% as ε increases. Although differential privacy initially increases iteration counts, DPFCM_GK demonstrates faster convergence and fewer iterations compared to DPFCM, with significant reductions (T= 23.08, 43.47, and 48.93; P<0.05). For the UCIML dataset, DPFCM_GK significantly reduces runtime compared to other models (DPFCM, LDP-SGD, LDP-Fed, LDP-FedSGD, MGM-DPL, LDP-FL) under the same privacy budget. The runtime reduction is statistically significant with T-values of (T = 21.08, 316.24, 102.35, 222.37, 162.23, 159.25; $P < 0.05$). DPFCM_GK still maintains excellent time efficiency when applied to the NLST and NSCLC datasets($P < 0.05$). For the LLCS dataset, For the LLCS dataset, the DPFCM_GK demonstrates significant improvement as the privacy budget increases, especially in low-budget scenarios, where the performance gap is most pronounced (T=4.20, 8.44, 10.92, 3.95, 7.16, 8.51, $P < 0.05$). These results confirm DPFCM_GK as a practical solution for medical data analysis, balancing accuracy, privacy, and efficiency.

Keywords Big data, Gaussian kernel function, Differential privacy, DPFCM_GK, Privacy-preserving, Privacy budget

As one of the most common malignant tumors worldwide, lung cancer necessitates an early diagnosis to ac qhieve higher cure rates and survival rates. However, the complex pathophysiological processes and elusive symptoms associated with lung cancer often pose enormous challenges to its early identification[1,2]. Conventional diagnostic methods rely heavily on the experience and expertise of physicians, but their accuracy and efficiency are limited, especially when dealing with large-scale medical data[3].

With the rapid development of medical imaging and bioinformatics technologies, leveraging machine learning (ML) and data mining techniques for lung cancer identification has become a hot research topic[4–6]. In the digital age, the collection, storage, and analysis of data have dramatically promoted the development across various sectors[7]. However, with the continuous growth and sharing of data, privacy protection has become an increasingly serious challenge, especially in the medical field. Safeguarding personal health data is of great significance as these data involve sensitive information such as medical history and genetic data[8,9]. Simultaneously, the privacy protection for medical data has become one of the significant challenges in lung cancer identification research. Conventional ML algorithms typically require training on centralized datasets.

Department of Respiratory and Critical Care Medicine, Affiliated Nanjing Gaochun People's Hospital, Jiangsu University, Nanjing 210000, Jiangsu, China. ✉email: zhaomingming10086@outlook.com

This implies that sensitive medical data need to be aggregated for processing, thus inducing potential risks for privacy breaches[10,11].

In recent years, privacy protection technologies such as differential privacy (DP) have received considerable attention as effective means of safeguarding sensitive information in data analysis and ML tasks[12,13]. DP provides a rigorous mathematical framework for quantifying and controlling privacy risks associated with the publication of statistical information derived from sensitive data. As an important privacy protection mechanism, DP has been widely applied in various scenarios, including healthcare, social networks, and financial services, to ensure the proper protection of individual privacy while allowing for the meaningful analysis and utilization of relevant data[14]. By introducing noise or perturbation to obfuscate data, DP techniques can provide robust protection for individual data privacy while maintaining data utility. This technology is favored not only in academia but also in policy-making and legal regulations. For example, some countries and regions have incorporated DP into data protection laws to ensure respect for individual privacy rights during data processing and sharing[15].

Kernel methods, particularly Gaussian kernel functions, have demonstrated remarkable effectiveness in various ML tasks, notably in clustering[16,17]. Gaussian kernels can be employed to flexibly map data into high-dimensional feature space with high efficiency, in which complex patterns can be captured and exploited for clustering purposes. The adaptability of Gaussian kernels arises from their ability to capture non-linear relationships inherent in the data, enabling robust and pronounced clustering performance[18,19]. In practice, Gaussian kernel-based approaches can be utilized to effectively handle complex data distributions, uncovering hidden structures that aid in accurate and insightful clustering outcomes[20]. The versatility and power of Gaussian kernels make them a cornerstone in modern ML. This contributes to an enhanced ability of ML to interpret and utilize intricate data patterns in such tasks as lung cancer identification and other challenging applications[21].

This study focuses on developing and optimizing privacy-preserving clustering methods tailored for medical data, addressing the critical challenge of safeguarding sensitive patient information while enabling effective data utilization. Privacy breaches in healthcare not only compromise individual rights but also hinder data sharing and collaborative research, both essential for advancing diagnostic methods like those for lung cancer. In this paper, a novel algorithm is proposed based on Gaussian kernel functions and DP techniques, integrated with the Fuzzy C-Means (FCM), for lung cancer identification[22–24]. By integrating advanced techniques such as differential privacy (DP) with clustering algorithms, this work aims to balance privacy protection with data utility. Specifically, the proposed approach ensures that sensitive medical information remains secure while preserving the accuracy and interpretability of clustering results, enabling reliable diagnostic insights.

The main contributions of this paper are elucidated as follows.

1. Propose a Gaussian kernel-based DP technique that can preserve model predictive performance while protecting individual privacy.
2. Construct a lung cancer identification algorithm based on DP techniques, which is capable of effectively mining potential features from medical data, thus achieving accurate cancer identification.
3. Conduct experimental validation on the proposed algorithm based on multiple public datasets, which demonstrates that this algorithm can achieve favorable lung cancer identification performance while protecting privacy.

Material and method

Data sources

In this study, the data are collected from three different platforms as experimental datasets, including the open-source dataset from the ML Repository of the University of California, Irvine (UCIML, http://archive.ics.uci.edu/dataset/62/)[25], the open dataset from the National Cancer Institute (NLST, https://cdas.cancer.gov/datasets/nlst/)[26], and the publicly available dataset released by Stanford University (NSCLC, https://wiki.cancerimagingarchive.net/display/Public/NSCLC+Radiogenomics)[27]. The experimental datasets are composed of patient demographics, including age distribution, gender, body mass index (BMI), and follow-up treatments sourced from systems maintained by the University of California, Irvine, the National Cancer Institute, and Stanford University. The collection of all data has been formally reviewed and approved by respective institutions, allowing access to these data for research purposes. In this retrospective study, the datasets from publicly available databases are utilized without any sensitive personal information, and hence no approval from ethics review boards is required. The effectiveness and portability of this algorithm are validated using the Lung Lesions and Cancer Set (LLCS), a lung cancer dataset collected by Gaochun Hospital, an affiliate of Jiangsu University. The LLCS comprises 256 case data, including 128 lung cancer cases and 128 other lung disease cases. The dataset includes various patient characteristics, such as height, weight, and imaging data. All LLCS data used in this study are collected in compliance with relevant guidelines and regulations (Ministry of Science and Technology of the People's Republic of China, Policy No. 2006398). This study has been approved by the Ethics Committee of Gaochun Hospital, Jiangsu University (No. K202300107). As a retrospective study, only the clinical data of patients are collected, without interfering with their treatment plans or imposing any physical risk. Besides, the informed consent from participants is not required for this retrospective study. However, the research team will take all necessary measures to ensure the confidentiality of patient information. This study will be conducted in accordance with good clinical practice and ethical standards set out in the Declaration of Helsinki of 1964 and its subsequent amendments.

DP technology

As a cutting-edge approach in the field of data privacy and security, DP aims to enable the sharing and analysis of sensitive data while safeguarding individual privacy. This revolutionary concept has received accumulating attention in recent years due to the increasing concern over data breaches and privacy violations in various

domains, such as healthcare, finance, and social media[28,29]. In essence, DP provides a mathematical framework for quantifying the privacy guarantees of a data-sharing mechanism. The fundamental idea is to add carefully calibrated noise to query responses or data records to mask sensitive information under the premise of ensuring the meaningful analysis of relevant data. The noise addition ensures that any single individual's data has a negligible impact on the overall results of the analysis, thereby preserving privacy[30].

One of the key advantages of DP is its rigorous formalization, which allows for precise privacy guarantees to be quantified and enforced. By carefully controlling the added amount of noise to the data, DP mechanisms can strike a balance between privacy and utility, ensuring that valuable insights can still be extracted from the data without compromising individuals' privacy[31]. The Notations and descriptions are shown in Table 1 below.

Specifically, if a dataset D and its corresponding random query function Q (query) are established, only a small probability difference between the output of the query function executed on any two adjacent datasets D and D' is required in DP.

For any adjacent datasets (D and D') and any output S, the following inequality can be obtained,

$$\Pr\left[Q\left(D\right) \in S\right] \leq \exp\left(\epsilon\right) * \Pr\left[Q\left(D'\right) \in S\right] \tag{1}$$

where, $\Pr$ represents the probability function; ϵ represents the privacy parameter of DP, also known as the privacy budget. The smaller the ϵ, the larger the added amount of noise. This indicates a lower probability of individual privacy disclosures and a higher degree of privacy protection.

As an important concept in DP, global sensitivity is used to measure the sensitivity of query results in different datasets and the maximum change in sensitivity functions caused by the deletion or addition of any data sample. Specifically, for a sensitive function f: D→R, where D represents the domain of the dataset and R represents the range of the sensitive function, the global sensitivity can be defined as follows,

$$\Delta f = max\left|f\left(D\right) - f\left(D'\right)\right| \; St.\, D, D' \rightarrow R \tag{2}$$

where, $|f\left(D\right) - f\left(D'\right)|$ represents the norm distance from $f\left(D\right)$ to $f\left(D'\right)$; the global sensitivity Δf represents the maximum sensitivity after noise is added to the dataset D. Here, the noise is generated randomly using the Laplacian mechanism, and the Laplacian noise density can be defined as follows,

$$Lap\left(x\right) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right) \tag{3}$$

where, b represents the scale parameter; μ represents the position parameter; x represents the position variable. In this function, when x= μ, $Lap\left(x|b\right) = \frac{1}{2b}$ is the largest, and both sides fall exponentially. The scale of the Laplace distribution, b, is determined by the sensitivity of the function and the desired privacy level. Sensitivity measures the maximum possible change in the function's output due to the addition or removal of a single record in the dataset. The scale is given by:

$$b = \frac{\Delta f}{\epsilon} \tag{4}$$

Notations	Descriptions	Notations	Descriptions
D	The dataset	c_j	The JTH cluster center
D'	Neighbor datasets of D	μ_{ij}	The membership of the sample x_i to the cluster center
S	Any set of outputs	ϵ	The convergence condition threshold
$\Pr$	The probability function	k || $*$ ||	Gaussian kernel functions
ϵ	The privacy budget	x'	The center of the kernel function
$f\left(*\right)$	The sensitive function	σ	The width parameter of the function
Δf	The global sensitivity	G()	The value of the Gaussian function
$Lap\left(x\right)$	Laplacian noise density	ω_i	The Gaussian weight of cluster i
B	The scale parameter	ϵ_i^t	The ϵ of the center point of cluster j during iteration t
X	The position variable	ACC	The accuracy
Q(*)	The random query function	PRE	The precision
J_m	The objective function	REC	Recall
M	The number of clusters of the cluster	ARI	Adjusted Rand index
N	The number of samples	RI	Rand index
C	The number of cluster centers	a	The number of samples divided into the same class
x_i	The i-th sample	d	The number of samples divided into different classes
||$*$||	The Euclidean norm		

Table 1. Notations and descriptions.

In the case of a constant Δf, the smaller the privacy budget, the larger the scale parameter b, and the greater the added noise.

For any function $f : \mathrm{D} \rightarrow R^k$, the Laplacian mechanism can be expressed as:

$$Q(x, f(*), \epsilon) = f(x) + (C_1, C_2, \ldots, C_k)$$

Where $C_i$is an independent equally distributed random variable subject to $C_i \sim \mathrm{Lap}\left(\frac{\Delta f}{\epsilon}\right)$ sampling, in fact, noise is added to the query result $f(x)$ of the database, the first output result is $f(x) + C_1$, and the k-th output result is $f(x) + C_k$.

In the theory of Laplace noise distribution, Δf, ϵ, μ, and b are the four most critical parameters. The proper configuration of these parameters directly influences the strength of privacy protection and the performance of clustering (as shown in Fig. 1). The position parameter (μ) represents the mean of the Laplace or Gaussian noise distribution, determining the center of the noise distribution. In this study, μ is set to 0 to ensure that the added noise does not introduce systematic bias to the clustering centers, thereby maintaining the randomness of privacy protection and the stability of model performance.

According to the properties of the Laplace noise distribution, the noise scale parameter b is proportional to Δf and inversely proportional to ϵ ($b = \Delta f/\epsilon$). As shown in Fig. 1B and D, and 1E, changes in Δf or ϵ independently lead to the same changes in the parameter b. To simplify the experimental process, Δf is fixed at 0.05 in this study. Instead, ϵ is dynamically adjusted to accommodate different scale parameters and sensitivity settings.

FCM

Fuzzy C-Means (FCM) is a clustering algorithm based on fuzzy theory. Compared with conventional K-means clustering algorithms, FCM allows data points to belong to multiple clusters at the same time by assigning a membership degree to each data point to describe its degree of belonging to each cluster[32,33]. FCM is widely used in fuzzy clustering and has achieved good results in image segmentation, pattern identification, data mining, and other fields.

Assuming that the dataset $D = \{x_1, x_2, x_3, \ldots, x_N\}$ is an n-dimensional sample data, FCM can obtain the cluster center by minimizing the objective function. The objective function is essentially the sum of the Euclidean distances from the points to the classes (sum of squares of errors),

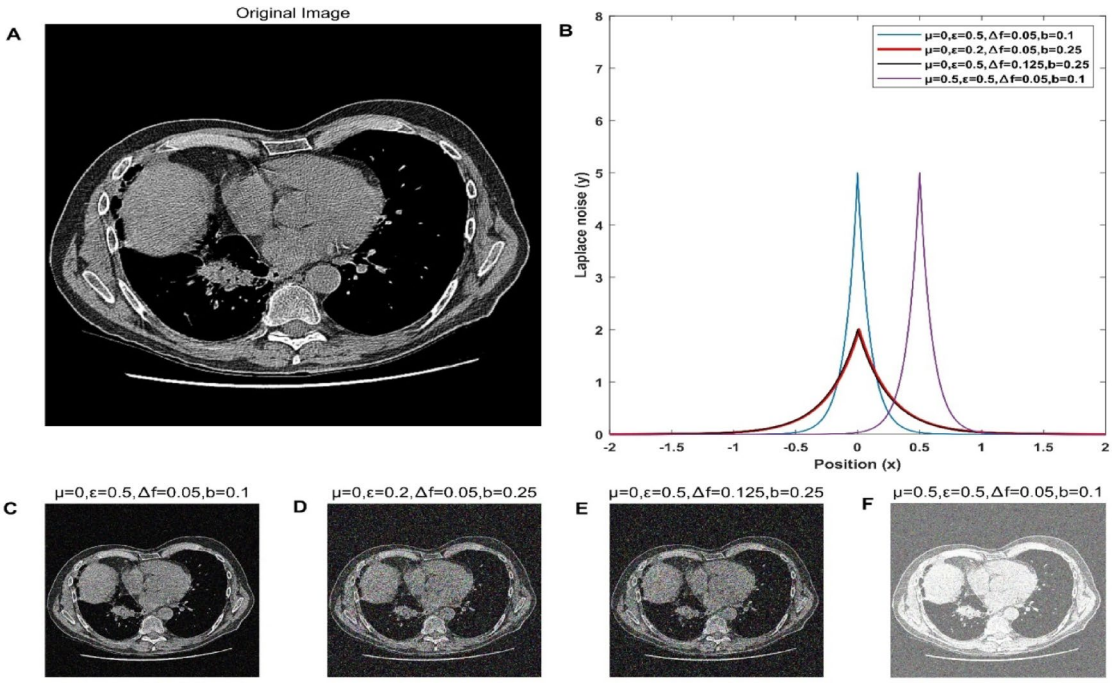

Fig. 1. The impact of Laplace noise settings with different parameters on image information. (**A**) The original lung CT image. (**B**) Laplacian noise distribution with different parameter settings. (**C**) The original lung cancer image after adding Laplacian noise with parameters $\mu = 0, \epsilon = 0.5, \Delta f = 0.05\ and\ b = 0.1$. (**D**) The original lung cancer image after adding Laplacian noise with parameters $\mu = 0, \epsilon = 0.2, \Delta f = 0.05\ and\ b = 0.25$. (**E**) The original lung cancer image after adding Laplacian noise with parameters $\mu = 0, \epsilon = 0.5, \Delta f = 0.125\ and\ b = 0.25$.(**F**) The original lung cancer image after adding Laplacian noise with parameters $\mu = 0.5, \epsilon = 0.5, \Delta f = 0.05\ and\ b = 0.1$..

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} \mu_{ij}^{m} || x_i - c_j ||^2 \qquad 1 \le m < \infty \tag{5}$$

where, m represents the number of clusters (number of classes) of the cluster; N represents the number of samples; C represents the number of cluster centers; x_i represents the i-th sample; $||*||$ represents the Euclidean norm; c_j represents the JTH cluster center, the same as the feature dimension of the sample; μ_{ij} represents the membership of the sample x_i to the cluster center (namely the probability of x_i belonging to c_j). Numerically, μ_{ij} and c_j can be calculated as follows,

$$\begin{cases} \mu_{ij} = \dfrac{1}{\sum_{k=1}^{C} \left(\frac{||x_i - c_j||}{||x_i - c_j||} \right)^{\frac{2}{m-1}}} \\ c_j = \dfrac{\sum_{i=1}^{N} \mu_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} \mu_{ij}^{m}} \end{cases} \tag{6}$$

where, μ_{ij} and c_j contain each other. Hence, in actual calculations, the assignment method is usually used for iterative calculations. When Eq. (7) is satisfied, the iteration ends.

$$\max_{i,j} \left\{ |\mu_{ij}^{t+1} - \mu_{ij}^{t}| \right\} < \epsilon \tag{7}$$

where, ϵ represents the convergence condition threshold, usually 0.01 or 0.001.

The function of combining DP with FCM based on Gaussian kernel function

Based on the analysis of the above privacy leakage problems, DP protection can be achieved by adding random noise satisfying the Laplace distribution to the center point of the clustering iterative process[34,35]. In each iteration, the same noise is added to the membership matrix and the cluster center point, which results in a large deviation of the cluster center point, thus eventually increasing the number of algorithm iterations and reducing the availability of relevant data. Therefore, the privacy budget al.location method combining DP with FCM based on the Gaussian kernel function (DPFCM_GK) can be constructed as follows.

The radial basis function (RBF) is a commonly used method for function approximation. RBF is based on a central point and determines the output value of the sample by calculating the distance between the input sample and the central point. In practice, Euclidean distance representation is often used in RBF[36–38]. As an important technique in ML, Gaussian kernel functions can be employed to clarify the relationship between Gaussian kernel functions and fuzzy sets, namely that Gaussian kernel functions can represent the relationship between objects,

$$\mathrm{RBF} = \mathrm{k}\,|| x - x' || = \exp\left(-\frac{|| x - x' ||^2}{2\sigma^2}\right) \tag{8}$$

where, x' represents the center of the kernel function; σ represents the width parameter of the function, and it is used to control the radial range of the function.

Let sensitive data x represent the data to be published, where $f(x) = x \in R^k$, i.e., D=x and $D' = x'$ be two neighboring datasets. So $\| f(D) - f(D') \| \le \Delta f$. For any output $o \in R^k$, the probability density of the Laplace mechanism is:

$$\begin{cases} Pr[f(D) = o] = \prod_{i=1}^{k} \frac{\epsilon}{2\Delta f} \exp\left(-\frac{\epsilon \left| f(D)_i - o_i \right|}{\Delta f} \right) \\ Pr[f(D') = o] = \prod_{i=1}^{k} \frac{\epsilon}{2\Delta f} \exp\left(-\frac{\epsilon \left| f(D')_i - o_i \right|}{\Delta f} \right) \end{cases} \tag{9}$$

The ratio of probabilities for D and D' is:

$$\frac{\Pr[f(D) = o]}{\Pr[f(D') = o]} = \frac{\prod_{i=1}^{k} \frac{\epsilon}{2\Delta f} \exp\left(-\frac{\epsilon \left| f(D)_i - o_i \right|}{\Delta f} \right)}{\prod_{i=1}^{k} \frac{\epsilon}{2\Delta f} \exp\left(-\frac{\epsilon \left| f(D')_i - o_i \right|}{\Delta f} \right)} = \prod_{i=1}^{k} \exp\left(\frac{\epsilon \left(\left| f(D')_i - o_i \right| - \left| f(D)_i - o_i \right| \right)}{\Delta f} \right) \tag{10}$$

Using the triangle inequality, we know:

$$\left| f(D')_i - o_i \right| - \left| f(D)_i - o_i \right| \le \left| f(D')_i - f(D)_i \right| \tag{11}$$

Summing over all *i*, we have:

$$\sum_{i=1}^{k} \left| f(D')_i - f(D)_i \right| \le \Delta f \tag{12}$$

Thus, the probability ratio becomes:

$$\frac{\Pr[f(D)=o]}{\Pr[f(D')=o]} = \prod_{i=1}^{k} \exp\left(\frac{\epsilon\left(|f(D')_i - o_i| - |f(D)_i - o_i|\right)}{\Delta f}\right) = \exp\left(\sum_{i=1}^{k} \frac{\epsilon\left(|f(D')_i - o_i| - |f(D)_i - o_i|\right)}{\Delta f}\right)$$
$$\leq exp\left(\frac{\epsilon}{\Delta f}\sum_{i=1}^{k}\left|f(D')_i - f(D)_i\right|\right) \leq exp\left(\frac{\epsilon}{\Delta f}\Delta f\right) = \exp(\epsilon) \tag{13}$$

For any measurable output set S, the same ratio bound holds by integrating the density function over S. This ensures:

$$\Pr[Q(D) \in S] \leq \exp(\epsilon) * \Pr\left[Q\left(D'\right) \in S\right] \tag{14}$$

Gaussian kernel functions have local properties and are suitable for allocating privacy budgets during cluster iterations. In each cluster, points farther away from the cluster center have smaller Gaussian kernel values, while those closer to the cluster center have larger Gaussian kernel values. The value of Gaussian kernel functions reflects the influence of the cluster center. A larger Gaussian value of the cluster center point suggests that the cluster points are more densely distributed around the cluster center, indicating a better clustering effect. Under this circumstance, a smaller privacy budget can be allocated to achieve a higher level of privacy protection[21].

Gaussian weights (ω_i) play a pivotal role in determining the allocation of privacy budgets to cluster centers. These algorithms leverage Gaussian functions to quantify the influence of data points on cluster centroids, with the weights reflecting the relative significance of each cluster in the overall clustering process. The algorithm defines $G(c_i)$ as the Gaussian function value for the i-th cluster center ci, computed based on the distances between the center ci and all data points x_j in the dataset. The Gaussian function $G(c_i)$ is formulated as,

$$G(c_i) = \sum_{j=1}^{N} \exp\left(-\frac{||x_j - x_i'||^2}{2\sigma^2}\right) \tag{15}$$

Where $||x_j - x_i'||^2$ represents the squared Euclidean distance between data point $x_j$and cluster center x_i'. σ is a parameter controlling the width of the Gaussian function, influencing how data points contribute to $G(c_i)$.

The Gaussian weight ω_i for the i-th cluster is then calculated as:

$$\omega_i = \frac{G(c_i)}{\sum_{j=1}^{k} G(c_j)} \tag{16}$$

Where k is the total number of clusters. This normalization ensures that ω_i represents the proportion of the total Gaussian function value contributed by cluster i, thereby quantifying its importance in the clustering process.

The allocation of privacy budgets ϵ_i^t to each cluster center c_i during iteration t is governed by the following formula:

$$\epsilon_i^t = \epsilon^t \cdot \frac{1 + \min_j \omega_j}{1 + \omega_i} \tag{17}$$

Where ϵ_i^t denotes the privacy budget allocated to the i-th cluster center in iteration t. ϵ^t represents the overall privacy budget available for the current iteration t. $\min_j \omega_j$ denotes the minimum Gaussian weight among all clusters, ensuring a baseline for privacy protection relative to the least influential cluster. ω_i is the Gaussian weight of the i-th cluster, determining the proportion of ϵ^t allocated based on its importance in the clustering process.

Impact of Gaussian Weights on Privacy Budgets:

- Higher ω_i values indicate that the *i*-th cluster center c_i has a greater influence on the clustering result. As a result, ϵ_i^t decreases because stricter privacy protections are required around influential centers to prevent privacy breaches.
- Lower ω_i values suggest that the corresponding cluster center c_i has less impact on the clustering outcome. Consequently, ϵ_i^tincreases, allowing for more relaxed privacy protection due to the reduced importance of data around these centers.

The formula $\epsilon_i^t = \epsilon^t \cdot \frac{1+\min_j \omega_j}{1+\omega_i}$ ensures mathematical rigor by scaling the global privacy budget ϵ^t inversely to 1+ ω_i, thereby adjusting for the impact of each cluster center's weight ω_i on the privacy guarantee. Utilizing $1 + \min_j \omega_j$ normalizes the privacy allocation across clusters, ensuring equitable distribution relative to the least contributing cluster, thereby maintaining fairness in privacy protection across all clusters.

The proposed model is designed to streamline operations through a structured workflow. The operational summary encompasses three key stages: input processing, core computation, and output generation, ensuring efficiency and accuracy. To clarify the interactions between these stages, Fig. 2 illustrates the operational summary of the proposed model.

In this study, a multimodal dataset comprising both image data and clinical feature data is collected. These image data consist of medical images, while these clinical feature data include physiological indicators, such as height, weight, and the number of structures. To leverage these data effectively, convolutional neural networks

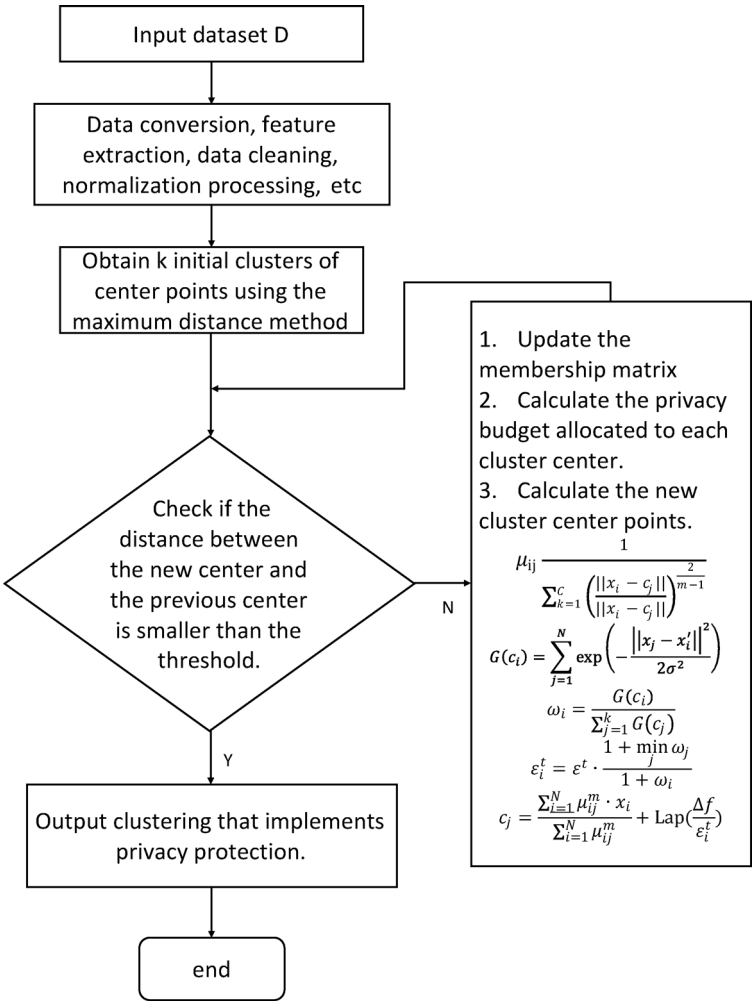

Fig. 2. Diagram depicting the operational summary of the proposed model.

(CNNs) are employed to extract features from the image data[39–41]. The DPFCM_GK algorithm was implemented in Python using libraries like NumPy, SciPy, and scikit-learn for numerical computations and clustering tasks. CNN-based feature extraction was performed with TensorFlow. Experiments were conducted on a workstation with an Intel Core i7 processor, 16 GB RAM, and an NVIDIA GeForce RTX 3080 GPU. Optimization techniques, including vectorized operations, batch processing, and early stopping, ensured computational efficiency. The CNN architecture for lung cancer image feature extraction consists of grayscale input images (256 × 256 pixels), four convolutional layers with 64 filters (3 × 3) and ReLU activation, max-pooling layers (2 × 2, stride 2), and fully connected layers (sizes 512 and 256). The network was trained using stochastic gradient descent (SGD) with momentum, a learning rate of 0.001, and dropout regularization (dropout rate 0.5) to reduce overfitting. Hyperparameters, including the learning rate, batch size, and dropout rate, were optimized through cross-validation to balance convergence and computational efficiency. These image features are combined with the clinical feature data to form the final feature set D. This integrated feature set D not only contains advanced representations of image data features but also incorporates information from clinical features, thus providing a comprehensive and diverse feature input for the subsequent data analysis and modeling.

Code availability

The custom code used in this study is available at Zenodo (DOI: https://doi.org/10.5281/zenodo.14789535) and can be accessed by readers for reproduction of the results and further analysis. The code is provided with no access restrictions and includes all necessary files for implementation. For additional details or questions, please contact the corresponding author.

Evaluation index

To verify the effectiveness of the joint clustering algorithm based on DP and Gaussian kernel functions, DPFCM_GK is employed to perform clustering procedures on four experimental data. The results are compared with those obtained based on the FCM and differential privacy fuzzy C-means clustering algorithm (DPFCM).

The accuracy (ACC), precision (PRE), recall (REC), F1 score (F1-score)[42], and adjusted Rand index (ARI)[43] are selected to evaluate the effectiveness of the algorithm,

$$\begin{cases} ACC = \frac{TP+TN}{TP+FN+FP+FN} \\ PRE = \frac{TP}{TP+FP} \\ REC = \frac{TP}{TP+FN} \\ F1\_Score = \frac{2*PRE*ReCall}{PRE+ReCall} \\ ARI = \frac{RI-E(RI)}{Max(RI)-E(RI)} \end{cases} \tag{18}$$

where, TP (true positive), FN (false negative), FP (false positive), and TN (true negative) represent the number of true classes, false negative classes, false positive classes, and true negative classes, respectively; RI represents the Rand index, which is equal to $\frac{2(a+b)}{n*(n+1)}$ numerically; a represents the number of samples divided into the same class; d represents the number of samples divided into different classes. The calculation results of each model were recorded using Excel software and subsequently analyzed with SPSS version 20.0. Continuous variables were expressed as mean ± standard deviation ($\bar{x} \pm s$). For repeated measurements across multiple datasets, repeated measures analysis of variance (ANOVA) was performed. Comparisons of continuous variables between two groups were conducted using paired t-tests. Categorical data were presented as percentages (%) and analyzed using chi-square (χ^2) tests to evaluate differences between groups. A p-value of less than 0.05 was considered statistically significant.

Result

Effectiveness analysis results

In this study, the clinical feature data of patients are modeled as the feature set due to the absence of image data in the UCIML dataset. The NLST dataset and NSCLC dataset contain image data, thus requiring the use of CNNs for feature extraction from the image data, followed by the combination with the clinical features to form the feature set. FCM, DPFCM, and DPFCM_GK are implemented using Python programming, and the clustering analysis is performed based on the UCIML, NLST, and NSCLC datasets, with the results shown in Fig. 3. As shown in Fig. 3, the privacy budget ε ranges from 0 to 5 with a step size of 0.01. The numbers of TP, FN, FP, and TN are recorded under different privacy budget ε values for the FCM, DPFCM, and DPFCM_GK models, and the values of ACC, PRE, ReCall, F1_Score, and ARI are also calculated.

It can be observed that the identification results of FCM based on the UCIML, NLST, and NSCLC datasets are independent of the privacy budget. However, the identification performance of DPFCM and DPFCM_GK is gradually improved with an increase in the privacy budget ε. Specifically, for the UCIML, NLST, and NSCLC datasets, DPFCM reaches the optimal clustering results when the privacy budgets are 2.65, 2.83, and 3.04, respectively; while DPFCM_GK reaches the optimal clustering results when the privacy budgets are 1.52, 1.24,

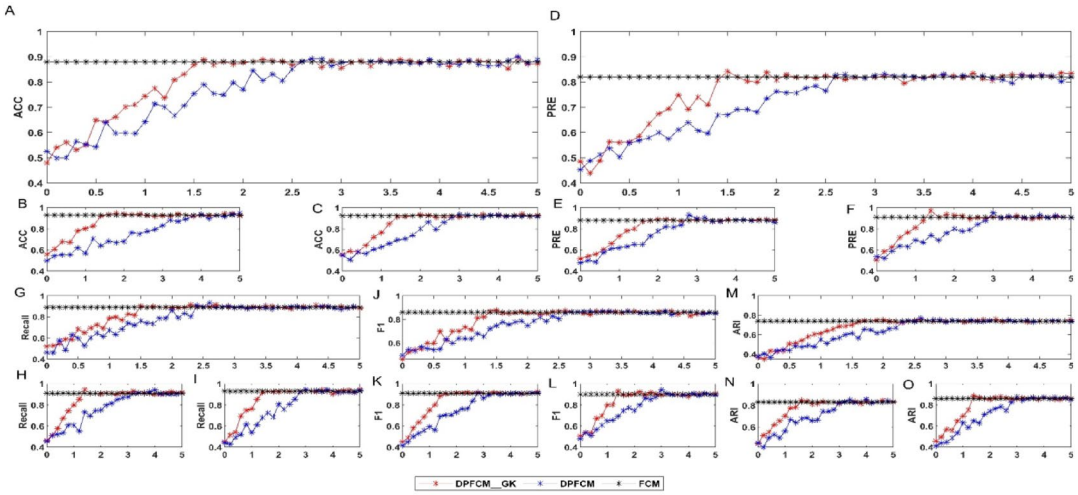

Fig. 3. Results of the effectiveness analysis of various algorithms based on the experimental data. Figure 3(**A**, **B**, **C**) illustrate the ACC calculation results based on the UCIML, NLST, and NSCLC datasets, respectively. Figure 3(**D**, **E**, **F**) illustrate the PRE calculation results based on the UCIML, NLST, and NSCLC datasets, respectively. Figure 3(**G**, **H**, **I**) illustrate the REC calculation results based on the UCIML, NLST, and NSCLC datasets, respectively. Figure 3(**J**, **K**, **L**) illustrate the F1-score calculation results based on the UCIML, NLST, and NSCLC datasets, respectively. Figure 3(**M**, **N**, **O**) illustrate the ARI calculation results based on the UCIML, NLST, and NSCLC datasets, respectively. The horizontal axis in all figures represents the privacy budget ε. The black line represents the identification results of FCM, the blue line represents the identification results of DPFCM, and the red line represents the identification results of DPFCM_GK.

and 2.32, respectively. As shown in Table 2, the comparison shows that DPFCM_GK outperforms DPFCM across privacy budgets ε. DPFCM_GK achieves better results in accuracy (ACC), precision (PRE), recall (ReCall), F1_Score, and adjusted rand index (ARI). Additionally, the P-value analysis confirms the differences are statistically significant, highlighting that DPFCM_GK strikes a better balance between privacy protection and model performance.

Comparison results of spatial and time complexity

To calculate the spatial complexity of each model, 12 different privacy budget values ε are adopted in this study, namely 0.01, 0.1, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5. Besides, a clustering analysis is performed on DPFCM and DPFCM_GK based on the UCIML, NLST, and NSCLC datasets, with the iteration counts of each algorithm recorded in Fig. 4.

Since noise can disrupt the original clustering convergence process, the iteration counts of algorithms with the implementation of DP protection will be higher than those without the implementation of DP protection. In other words, when the privacy budget is small, the iteration counts of DPFCM_GK and DPFCM are higher than those of FCM. As the privacy budget gradually increases, the added random noise decreases gradually, resulting in a reduction in the average iteration counts of the two DP protection algorithms, approaching those of FCM. Meanwhile, DPFCM_GK exhibits a faster convergence trend and significantly faster convergence speed, exhibiting statistical significance ($T = 23.08, 43.47$, and 48.93; $P < 0.05$).

Furthermore, this study conducted an in-depth analysis of the time complexity of the proposed DPFCM_GK method and compared it with several existing privacy-preserving methods, including LDP-SGD, LDP-Fed, LDP-FedSGD, MGM-DPL, and Ldp-fl[44–48]. The iteration time of FCM, DPFCM, DPFCM_GK, LDP-SGD, LDP-Fed, LDP-FedSGD, MGM-DPL, and Ldp-fl are also calculated for each privacy budget state, as shown in Fig. 5. For the UCIML dataset, under the same privacy budget, the DPFCM_GK model demonstrates significantly lower time complexity, greatly reducing the runtime compared to models like DPFCM, LDP-SGD, LDP-Fed, LDP-

Data set	Metric	Model	Privacy budget ε						P Value
			0.01	1	2	3	4	5	
UCIML	ACC	DPFCM_GK	0.568	0.779	0.870	0.875	0.891	0.885	3.95E-4
		DPFCM	0.535	0.632	0.844	0.864	0.863	0.889	
	PRE	DPFCM_GK	0.504	0.686	0.817	0.822	0.809	0.827	5.07E-06
		DPFCM	0.486	0.592	0.674	0.815	0.831	0.815	
	ReCall	DPFCM_GK	0.547	0.726	0.856	0.898	0.886	0.881	4.59E-05
		DPFCM	0.525	0.674	0.767	0.881	0.889	0.869	
	F1_Score	DPFCM_GK	0.470	0.756	0.837	0.843	0.868	0.856	5.68E-05
		DPFCM	0.565	0.654	0.792	0.874	0.849	0.861	
	ARI	DPFCM_GK	0.380	0.604	0.735	0.743	0.754	0.735	3.98E-06
		DPFCM	0.353	0.492	0.693	0.733	0.741	0.739	
UCIML	ACC	DPFCM_GK	0.578	0.832	0.914	0.907	0.938	0.931	3.96E-06
		DPFCM	0.543	0.643	0.680	0.829	0.902	0.920	
	PRE	DPFCM_GK	0.516	0.720	0.894	0.888	0.892	0.869	1.51E-04
		DPFCM	0.525	0.615	0.749	0.890	0.859	0.892	
	ReCall	DPFCM_GK	0.562	0.801	0.908	0.911	0.901	0.898	4.65E-03
		DPFCM	0.499	0.593	0.830	0.902	0.922	0.903	
	F1_Score	DPFCM_GK	0.530	0.778	0.907	0.909	0.913	0.917	5.46E-04
		DPFCM	0.505	0.586	0.791	0.928	0.912	0.914	
	ARI	DPFCM_GK	0.483	0.730	0.849	0.832	0.843	0.818	2.90E-04
		DPFCM	0.420	0.569	0.711	0.837	0.831	0.830	
UCIML	ACC	DPFCM_GK	0.594	0.737	0.910	0.932	0.918	0.918	1.17E-03
		DPFCM	0.513	0.617	0.810	0.999	0.916	0.928	
	PRE	DPFCM_GK	0.568	0.797	0.901	0.925	0.913	0.914	5.43E-04
		DPFCM	0.513	0.631	0.699	0.905	0.908	0.911	
	ReCall	DPFCM_GK	0.516	0.772	0.944	0.919	0.942	0.920	1.89E-04
		DPFCM	0.472	0.594	0.777	0.929	0.935	0.930	
	F1_Score	DPFCM_GK	0.574	0.825	0.893	0.893	0.912	0.911	4.56E-04
		DPFCM	0.540	0.629	0.768	0.897	0.898	0.903	
	ARI	DPFCM_GK	0.46	0.647	0.860	0.860	0.862	0.865	2.74E-04
		DPFCM	0.475	0.582	0.703	0.851	0.873	0.852	

Table 2. Results of the effectiveness analysis of DPFCM and DPFCM_GK. Note: In the table, only results for privacy budget ε = 0.01, 1, 2, 3, 4, 5 are included. The P Value represents the results of a paired t-test between DPFCM and DPFCM_GK.

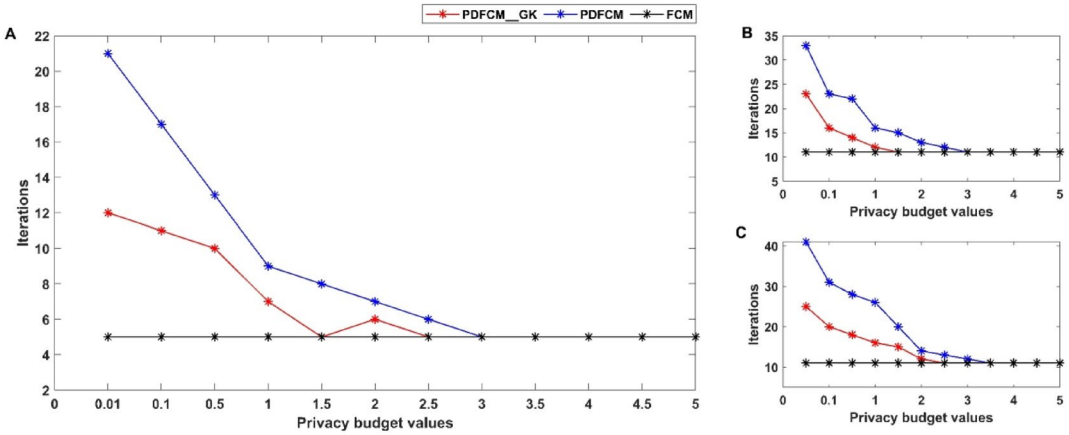

Fig. 4. Analysis results of clustering iteration times for various algorithms based on the experimental data. Figure 4(**A**, **B**, **C**) illustrate the calculation results of clustering iteration counts for each algorithm based on the UCIML, NLST, and NSCLC datasets, respectively.

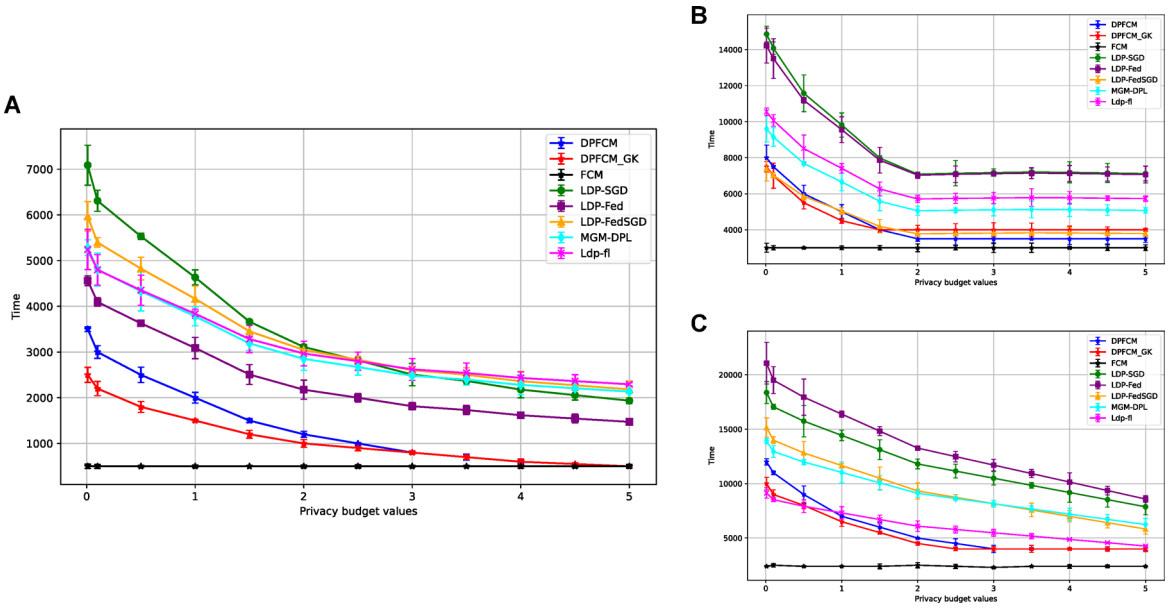

Fig. 5. Analysis results of clustering running time (ms) of various algorithms based on the experimental data. Figure 5(**A**, **B**, **C**) illustrate the calculation results of clustering iteration time (ms) for each algorithm based on the UCIML, NLST, and NSCLC datasets, respectively.

FedSGD, MGM-DPL, and LDP-FL. Even at higher privacy budget values, DPFCM_GK continues to perform efficiently. The reduction in runtime is statistically significant, with T-values of (T = 21.08, 316.24, 102.35, 222.37, 162.23 and 159.25, $P < 0.05$). For the NLST and NSCLC datasets, the DPFCM_GK algorithm still maintains the lowest runtime across all models(T = 26.88, 217.66, 132.19, 267.33, 182.12 and 144.13, $P < 0.0$ 182.12 and 144.13, $P < 0.05$; T = 31.78, 267.89, 156.71, 265.38, 191.22 and 124.56. $P < 0.05$), demonstrating its superior time efficiency.

Algorithm availability verification

To validate the utility of the data, a classification analysis is performed on the lung cancer dataset collected from our hospital using FCM, DPFCM, and DPFCM_GK to verify the robustness and portability of these algorithms. The privacy budget ε is set to 0.01, 0.1, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5, respectively. The numbers of TP, FN, FP, and TN identified by FCM, DPFCM, and DPFCM_GK under various conditions are also recorded, and their identification results are simultaneously calculated, as shown in Fig. 6.

For the lung cancer dataset collected by our hospital, as ε increases, the clustering results of the DPFCM-GF and DPFCM algorithms gradually converge to those of the FCM algorithm, achieving an accuracy rate of 94.5%. With in the range of $0.5 \leq \varepsilon \leq 2.5$, the clustering performance of the DPFCM-GF algorithm is superior

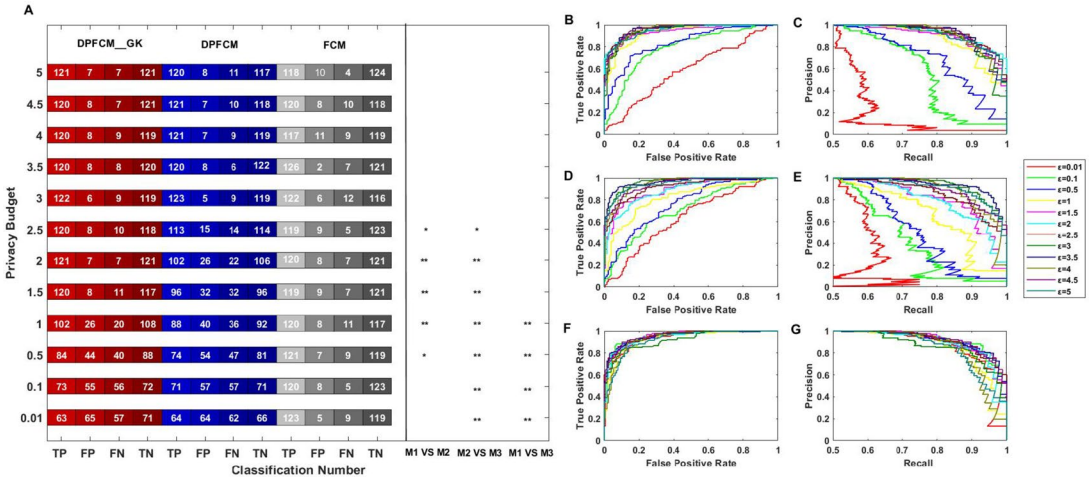

Fig. 6. Performance evaluation and availability verification results of DPFCM_GK, DPFCM, and FCM using LLCS. (**A**) Comparison of identified cases for DPFCM_GK, DPFCM, and FCM. The numbers inside the colored blocks represent the count of identified cases corresponding to the horizontal axis. Red blocks indicate the results of DPFCM_GK, blue blocks indicate the results of DPFCM, and gray blocks indicate the results of FCM. M1, M2, and M3 denote DPFCM_GK, DPFCM, and FCM, respectively. The symbol "*" represents a statistically significant difference between two algorithms with a p-value less than 0.05, while "**" represents a p-value less than 0.01. (**B**, **D**, **F**) ROC curves for DPFCM_GK, DPFCM, and FCM, respectively, illustrating their performance under LLCS classification. (**C**, **E**, **G**) PR curves for DPFCM_GK, DPFCM, and FCM, respectively, demonstrating their precision-recall relationships under LLCS classification.

to that of the DPFCM algorithm, with statistically significant differences ($\chi^2 = 4.54, 9.68, 29.12, 21.21$, and 4.34; $P < 0.05$). The ROC and PR curves of the LLCS dataset were drawn. As the privacy budget ε increases, the Area Under the Curve (AUC) values predicted by the DPFCM and DPFCM-GF algorithms for the ROC curve of the LLCS dataset converge to 0.84. The DPFCM-GF algorithm demonstrates faster convergence and better recognition performance compared to DPFCM.

By setting different privacy budgets, the study evaluated the classification accuracy and data loss rate of each model on the LLCS dataset, providing a comprehensive comparison of the performance of different algorithms under varying privacy budget conditions. Furthermore, with the privacy budget fixed at 5, this study analyzed the relationship between the number of iterations and the data loss rate for each model. The results, presented in Fig. 7, clearly demonstrate the advantages of the DPFCM_GK method in balancing convergence speed and data loss rate. The DPFCM_GK method consistently achieves higher accuracy compared to other methods (DPFCM, LDP-SGD, LDP-Fed, LDP-FedSGD, MGM-DPL, and Ldp-fl) across all privacy budget levels. The DPFCM_GK method achieves the lowest misclassification rate among all models, particularly when the privacy budget is small. As the privacy budget increases, the misclassification rate decreases for all methods, but DPFCM_GK maintains its advantage (T =4.20, 8.44, 10.92, 3.95, 7.16, 8.51, $P < 0.05$). The DPFCM_GK method converges faster to a lower misclassification rate compared to other methods, highlighting its efficiency in achieving better results within fewer iterations. When the privacy budget ε is set to 5, the confusion matrices of the DPFCM, DPFCM_GK, LDP-Fed, LDP-FL, LDP-SGD, MGM-DPFL, and MGM-FedSGD models are evaluated using the LLCS dataset (as shown in Fig. 7D and J).

Discussion

Medical data refer to various information related to patients' health, and they are collected, stored, and processed in the healthcare field, including personal physical conditions, medical records, diagnostic results, treatment plans, and prescription medications[49]. These data flow among medical institutions, healthcare providers, researchers, and insurance companies for disease diagnosis, treatment planning, medical research, and healthcare service management. However, the privacy breach of medical data could pose serious risks[50–52]. Firstly, the leakage of sensitive health information may lead to infringement of personal privacy rights, even exposing individuals to risks like identity theft and financial loss. Secondly, the leakage of medical data may weaken individuals' trust in the healthcare system and reduce their willingness to seek medical care and share health information, thereby affecting the quality and effectiveness of healthcare services. Additionally, the unauthorized use of leaked medical data could be exploited by malicious actors for cyberattacks, resulting in the paralysis of healthcare systems and posing serious threats to patient safety. Therefore, there is an urgent demand for realizing the sharing and fusion of data under the premise of protecting the privacy of medical data[53,54].

To optimize or solve the problems associated with medical data security and privacy protection, data encryption and data protection technologies have been proposed and explored by many researchers[55,56]. DP is a data processing framework that protects personal privacy information by adding noise to the original data, making it difficult for attackers to infer sensitive personal information from the noise. The application of DP

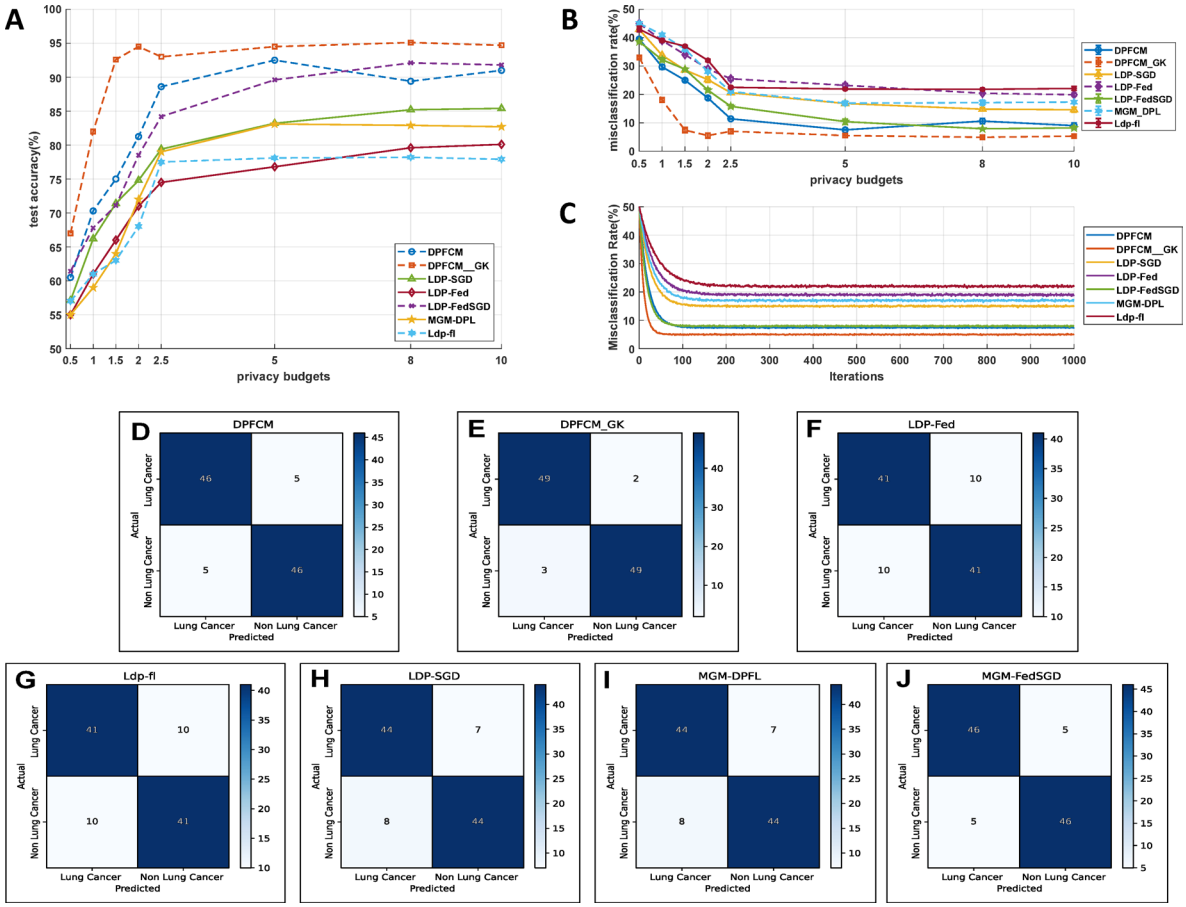

Fig. 7. Performance evaluation of DPFCM_GK and other methods under varying privacy budgets using LLCS. (**A**) The relationship between the test accuracy (%) of various models and privacy budget values. (**B**) The relationship between the misclassification rate (%) of various models and privacy budget values. (**C**) The relationship between the misclassification rate (%) of various models and the number of iterations, with the privacy budget fixed at 5. (**D-J**) When the privacy budget ε is set to 5, the confusion matrices of the DPFCM, DPFCM_GK, LDP-Fed, LDP-FL, LDP-SGD, MGM-DPFL, and MGM-FedSGD models are evaluated using the LLCS dataset.

in medical data security has exhibited great potential, which can effectively protect the privacy of medical data while allowing for data sharing and analyses.

With the advancement of deep learning (DL) algorithms, particularly convolutional neural networks (CNNs), significant progress has been made in applying DL-based models to computer-aided diagnosis (CAD) systems for lung tumor detection. These models aim to improve diagnostic accuracy, reduce false positive rates, and enhance execution efficiency (Table 3)[57,58]. Much like feature-based CAD systems, DL-based systems typically follow a three-step workflow: nodule detection and segmentation, feature extraction, and clinical judgment inference[59]. However, unlike traditional methods that rely on manually crafted features, DL-based CAD systems can automatically learn and extract intrinsic features from nodules[60,61], as well as model their 3D shapes. This automatic feature extraction capability allows for more accurate and comprehensive analysis of lung nodules. For instance, Ciompi et al.[62] developed a model based on OverFeat[63,64], which extracts 2D-view feature vectors from CT scans in three planes: axial, coronal, and sagittal. These vectors enable a multi-view analysis of the nodule, capturing essential features for classification. Recent CNN models have further advanced this approach by providing a global, holistic view of nodules for more detailed feature characterization from CT images. Buty et al.[59] proposed a complementary CNN model that incorporates both shape and appearance features for lung nodule classification. In their approach, a spherical harmonic model[60] is used for nodule segmentation, which yields shape-based descriptors of the nodule. Simultaneously, a deep convolutional neural network (DCNN)[63] is employed to extract texture and intensity features, referred to as appearance features. These two types of features are then combined to enhance the downstream classification process, enabling the model to effectively distinguish between benign and malignant nodules. Similarly, Venkadesh et al.[66] implemented an ensemble approach by combining two different CNN models: a 2D-ResNet50-based model[67] and a 3D-Inception-V1 model[68]. Each of these models specializes in extracting specific features from pulmonary nodules. The 2D-ResNet50 model focuses on extracting 2D features, while the 3D-Inception-V1 model captures 3D information. By concatenating the features extracted by these models, the ensemble model achieves superior performance in identifying

Publication	Feature extraction	Classification model	Sample size	Data type	AUC	ACC	Validation method	Feature selection/input	Highlight/advantage	Shortcoming
Mohammad[57]	NA	LR	300	CT images	0.706–0.932	NA	Hold-out	Clinical factors + nodule characteristics	The classifier can perform equivalently as human observers for malignant and benign classification	The performance heavily relies on nodule size as the discriminator, and is not robust in small nodules
Armato[58]	NA	LDA	326	MALDI	NA	0.991	Hold-out	Mass spectra from ROIs of MALDI	The model maintains high accuracy on FFPE biopsies	The performance relies on the quality of the MALDI stratification
Buty[59]	Spherical harmonics[66]; DCNN[63]	RF	1018	CT images	NA	0.793–0.824	10-fold cross-validation	CT imaging patches + radiologists' binary nodule	The model reaches higher predictive accuracy by integrating shape and appearance nodule imaging features	No benchmarking comparisons were used in the study
Hussein[60]	3D CNN-based multi-task model	3D CNN-based multi-task model	1018	CT images	NA	0.9126	10-fold cross-validation	3D CT volume feature	The model achieves higher accuracy than other benchmarked models	The ground truth scores defined by radiologists for the benchmark might be arbitrary
Khosravan[61]	3D CNN-based multi-task model[65]	3D CNN-based multi-task model	6960	CT images	NA	0.97	10-fold cross-validation	3D CT volume feature	Algorithms helps to accurately extract potential attentional regions	Segmentation might fail if the ROIs are outside the lung regions
Ciompi[62]	OverFeat[64]	SVM; RF	1729	CT images	0.868	NA	10-fold cross-validation	3D CT volume feature, nodule position coordinate et al.	This is the first study attempting to classify whether the diagnosed nodule is benign or malignant	Many nodules could not be located on the CT images
Venkadesh[66]	2D-ResNet50-based[67,68]	An ensemble model based on two CNN models	16,429	CT images	0.86–0.96	NA	10-fold cross-validation	3D CT volume feature and nodule coordinates	The model achieves higher AUC than other benchmarked models	The model requires specifying the position of the nodule, but many nodules are unable to be located on the CT images
Coudray et al.[69]	Multi-task CNN model based on Inception-V3	Multi-task CNN model based on Inception-V3	1634	Histological images	0.733–0.856	NA	Hold-out	Transformed 512 × 512-pixel tiles	The model can predict whether a given tissue has somatic mutations in genesSTK11,EGFR, et al.	The accuracy of the gene mutation prediction is not very high
Lin et al.[70]	DCGAN + AlexNet	DCGAN + AlexNet	22,489	CT images	NA	0.9986	Hold-out	Initial + synthetic CT images	The model uses GAN to generate synthetic lung cancer images to reduce overfitting	No benchmarking comparisons were used
Ren et al.[71]	DCGAN + VGG-DF	DCGAN + VGG-DF	15,000	Histopathological images	NA	0.9984	Hold-out	Initial + synthetic histopathological images	The model uses GAN to generate synthetic lung cancer images and reduce overfitting	The dimension is not sufficient for biomedical domain

Table 3. Publications relevant to ML on detection and diagnosis.

malignant nodules of varying sizes from raw CT images. The ensemble CNN approach offers several advantages, particularly in its ability to accurately classify nodules of diverse shapes, sizes, and textures. Leveraging the capabilities of state-of-the-art CNNs, these models can generate high-quality features that significantly improve diagnostic outcomes. Once these features are extracted, clinical judgment inference can be performed using various machine learning (ML) techniques, including logistic regression (LR), random forest (RF), support vector machine (SVM), and neural networks (NNs).

Another study utilized the Inception-V3 network to classify whether the tissue was LUAD, LUSC, or normal from H&E-stained histopathology whole-slide images[69]. A highlight of this study is that the model can also predict whether a given tissue has somatic mutations in several lung cancer driver genes, including STK11, EGFR, FAT1, SETBP1, KRAS, and TP53. Lin et al.[70] used a two-step model — a DCGAN to generate synthetic lung cancer images and an AlexNet for lung cancer classification using both original and synthetic datasets. Similar work was also done by Ren and colleagues[71]. They also used DCGAN for data augmentation. To improve

performance, they then designed a regularization-enhanced transfer learning model called VGG-DF for data discrimination to prevent overfitting problems with pre-trained model auto-selection.

Although the aforementioned studies have made significant advancements in the identification of lung nodules and lung cancer using artificial intelligence (AI) and computer-aided diagnosis (CAD) systems, the issue of patient privacy protection has received limited attention in these works. This lack of focus on privacy is particularly concerning, given that AI-based systems often require access to sensitive medical data, such as patient CT scans and health records, to train and validate their models. Most existing methods prioritize improving diagnostic accuracy, reducing false positives, and optimizing computational efficiency, but they largely overlook the critical ethical and legal implications of handling patient data. Traditional data anonymization techniques, such as removing identifiable information, are no longer sufficient to guarantee privacy in the era of AI. Advanced re-identification techniques can potentially reconstruct patient identities from de-identified medical images or data patterns, especially when combined with external data sources. This highlights the critical need for robust privacy-preserving mechanisms in AI-driven CAD systems. However, privacy-preserving approaches, such as federated learning and differential privacy, have been rarely applied in the context of lung cancer detection or CAD systems. This oversight leaves a significant gap in the current research landscape.

In the DP research, clustering algorithms are employed to classify data into sets with similar characteristics, which can be used to discover hidden patterns and correlations in the data. However, conventional DP clustering algorithms often face a conflict between privacy protection and data utility. On one hand, adding noise to protect privacy may introduce risks related to privacy leakage, leading to inaccurate clustering results[72]. On the other hand, reducing noise to improve the accuracy of clustering results may compromise the effectiveness of privacy protection.

In this study, a clustering algorithm is proposed by combining DP with the Gaussian kernel function. The Gaussian kernel function is a commonly used kernel function that maps data into high-dimensional space for better clustering and classification. The controlled noise allows for more precise clustering results while protecting data privacy. In this study, the DP technique based on the Gaussian kernel function provides an effective means of securing medical data. Through noise addition, DP can protect sensitive information while maintaining the accuracy of clustering results. Based on four datasets, such clustering metrics as ACC, PRE, ReCall, F1_Score, and ARI gradually align with those of FCM as the privacy budget increases, indicating that the Gaussian kernel function can perform the clustering analysis reliably while preserving data privacy. Additionally, the proposed DPFCM_GK converges rapidly with an increase in the privacy budget. Compared with conventional DP algorithms, DPFCM_GK exhibits faster convergence, fewer iterations, and lower time complexity. Moreover, DPFCM_GK is verified using a lung cancer dataset collected from our hospital. The results indicate that as the privacy budget ε increases, DPFCM_GK can effectively perform lung cancer identification and classification. This validates that this algorithm can also effectively limit the risk of privacy leakage under the premise of ensuring data utility.

The success of DPFCM_GK lies in its innovative integration of differential privacy mechanisms with the robust fuzzy clustering framework. By adding carefully calibrated noise to the data or computations, DPFCM_GK effectively protects sensitive information while maintaining high clustering accuracy. Its fuzzy clustering approach assigns degrees of membership to each data point, enhancing tolerance to noise and uncertainty, which is crucial in privacy-preserving scenarios. The algorithm is designed to balance privacy and utility by optimizing parameters such as the privacy budget ε, ensuring minimal performance degradation while achieving strong privacy guarantees. Furthermore, the mathematical enhancements to the traditional FCM model, such as adjusting the calculation of membership degrees and cluster centers with privacy-preserving constraints, allow DPFCM_GK to perform well under differential privacy constraints. This combination of privacy-preserving design, robustness to noise, and optimized performance makes DPFCM_GK highly effective for sensitive data clustering in practical applications.

In this study, the effect of the privacy budget ε on the convergence behavior of the DPFCM_GK algorithm was analyzed. When ε decreases, representing stricter privacy constraints, the Gaussian mechanism introduces more significant noise into the system. This noise interferes with the gradient descent process by distorting the descent direction, which slows the convergence of the algorithm and increases fluctuations in the stability of the clustering results. On the other hand, higher values of ε, corresponding to weaker privacy constraints, reduce the noise added by the mechanism. This allows the algorithm to achieve faster convergence and enhances the consistency and reliability of clustering outcomes. This relationship highlights the inherent trade-off between ensuring rigorous privacy protection and maintaining optimal model performance. Stricter privacy measures safeguard data confidentiality but at the cost of computational efficiency and clustering accuracy. Conversely, relaxing privacy constraints benefits algorithm stability and convergence but reduces the level of privacy protection.

However, the limitations of the DPFCM_GK approach are crucial to providing a balanced view of its performance. Specifically: While the joint clustering algorithm demonstrates effectiveness with small to medium-sized datasets, it faces computational challenges when applied to larger datasets. As the dataset size increases, the time complexity of the algorithm rises, which may hinder its scalability for real-world applications involving large volumes of data. The differential privacy mechanism integrated into the method introduces a trade-off between ensuring data privacy and maintaining data utility. The noise added to safeguard privacy may reduce the accuracy of the clustering results, particularly when working with sensitive datasets. This trade-off should be considered when implementing the approach in practical, privacy-sensitive domains. The algorithm was primarily evaluated on a specific dataset. Although the results were promising, further exploration is needed to evaluate its adaptability to diverse data types or datasets from different fields. The algorithm's performance may vary when applied to data with different structures or characteristics, suggesting that more testing is needed to ensure its generalizability.

While the results highlight the effectiveness of our method, we recognize the importance of integrating more recent advancements in the field to further strengthen the robustness of our approach. One such advancement is the introduction of quantum machine learning algorithms, as discussed by Rishabh Gupta et al. in "Quantum Machine Learning Driven Malicious User Prediction for Cloud Network Communications." This paper introduces an innovative model for predicting malicious users in cloud network communications, leveraging quantum machine learning to efficiently identify and predict malicious users. This approach enhances system security by proactively detecting malicious entities before a data breach occurs[73]. Furthermore, in their subsequent research, they developed the DT-PPM model, which employs techniques such as data partitioning, noise injection, and classification to safeguard medical data in cloud environments. This model offers a secure way to share outsourced data from pathology centers while ensuring privacy protection and significantly reducing computation time[74].

Their research shares similarities with our approach, particularly in their use of noise injection to protect data privacy. However, their focus on multi-center data, especially data from various pathology centers, presents challenges that are not yet addressed in our current study. While we focus on adding Laplace noise to medical data using a Gaussian kernel, we have not yet explored the complexities of privacy protection in a multi-center environment. Future work will aim to incorporate these advanced techniques to tackle privacy issues in multi-center data environments. Specifically, we plan to extend our method to include mechanisms that protect data from multiple pathology centers, where privacy concerns are more pronounced due to the involvement of diverse data sources. This will require adapting our Gaussian kernel-based approach to effectively handle data partitioning and multi-center noise injection, building on the foundation set by the reference model.

Conclusion

In conclusion, the proposed DPFCM_GK offers a robust solution for cancer identification, ensuring both data accuracy and privacy. This algorithm exhibits faster convergence, fewer iterations, and lower time complexity compared with conventional DP algorithms. Therefore, it is considered an appealing choice for real-world applications in medical data analyses. Overall, it represents a significant step forward in securing sensitive medical data, especially in the field of cancer identification. Its applicability to other medical datasets and its potential for broader healthcare applications may be explored in further studies.

Data availability

This study utilized data from three different platforms as experimental datasets. They are: the open-source dataset from the Machine Learning Repository of the University of California, Irvine (UCIML, http://archive.ics.uci.edu/dataset/62/), the open dataset from the National Cancer Institute (NLST, https://cdas.cancer.gov/datasets/nlst/), and the publicly available dataset released by Stanford University (NSCLC, https://wiki.cancerimagingarchive.net/display/Public/NSCLC+Radiogenomics). All data and materials are fully available by Zhao Mingming (zhaomingming10086@outlook.com). He will provide necessary information and data access permissions.

Received: 14 May 2024; Accepted: 8 May 2025
Published online: 16 May 2025

References

1. BadeBC & Dela CruzCS Lung Cancer 2020: epidemiology, etiology, and prevention. *Clin. Chest Med.* **41** (1), 1–24. https://doi.org/10.1016/j.ccm.2019.10.001 (2020).
2. Nasim, F., Sabath, B. F. & Eapen, G. A. Lung Cancer. *Med. Clin. North. Am.* **103** (3), 463–473. https://doi.org/10.1016/j.mcna.2018.12.006 (2019).
3. Lockery, J. E. et al. Optimising medication data collection in a large-scale clinical trial. *PLoS One*. **14** (12), e0226868. https://doi.org/10.1371/journal.pone.0226868 (2019). Published 2019 Dec 27.
4. Greener, J. G., Kandathil, S. M., Moffat, L. & Jones, D. T. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell. Biol.* **23** (1), 40–55. https://doi.org/10.1038/s41580-021-00407-0 (2022).
5. Deo, R. C. Machine learning in medicine. *Circulation* **132** (20), 1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593 (2015).
6. Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F. & Campbell, J. P. Introduction to machine learning, neural networks, and deep learning. *Transl Vis. Sci. Technol.* **9** (2), 14. https://doi.org/10.1167/tvst.9.2.14 (2020). Published 2020 Feb 27.
7. Kayaalp, M. Patient privacy in the era of big data. *Balkan Med. J.* **35** (1), 8–17. https://doi.org/10.4274/balkanmedj.2017.0966 (2018).
8. Zhang, S., Li, M., Liang, W., Sandor, V. K. A. & Li, X. A survey of Dummy-Based location privacy protection techniques for location-Based services. *Sens. (Basel)*. **22** (16), 6141. https://doi.org/10.3390/s22166141 (2022). Published 2022 Aug 17.
9. Wang, H. et al. Privacy protection generalization with adversarial fusion. *Math. Biosci. Eng.* **19** (7), 7314–7336. https://doi.org/10.3934/mbe.2022345 (2022).
10. Moore, W. & Frye, S. Review of HIPAA, part 1: history, protected health information, and privacy and security rules. *J. Nucl. Med. Technol.* **47** (4), 269–272. https://doi.org/10.2967/jnmt.119.227819 (2019).
11. Ns Chan, D., Choi, K. C., Hy To, M., Kn Ha, S. & Cc Ling, G. Patient privacy protection among university nursing students: A cross-sectional study. *Nurs. Ethics*. **29** (5), 1280–1292. https://doi.org/10.1177/09697330221085777 (2022).
12. Ficek, J., Wang, W., Chen, H., Dagne, G. & Daley, E. Differential privacy in health research: A scoping review. *J. Am. Med. Inf. Assoc.* **28** (10), 2269–2276. https://doi.org/10.1093/jamia/ocab135 (2021).
13. Müllner, P., Lex, E., Schedl, M. & Kowald, D. Differential privacy in collaborative filtering recommender systems: a review. *Front. Big Data*. **6**, 1249997. https://doi.org/10.3389/fdata.2023.1249997 (2023). Published 2023 Oct 12.
14. Bi, X., Shen, X. & Distribution-Invariant Differential, P. *J. Econom* ;**235**(2):444–453. doi:https://doi.org/10.1016/j.jeconom.2022.05.004 (2023).
15. Liu, W., Zhang, Y., Yang, H. & Meng, Q. A Survey on Differential Privacy for Medical Data Analysis. *Ann. Data Sci. Published online June* **10**, doi:https://doi.org/10.1007/s40745-023-00475-3 (2023).
16. Schaid, D. J. Genomic similarity and kernel methods II: methods for genomic information. *Hum. Hered.* **70** (2), 132–140. https://doi.org/10.1159/000312643 (2010).

17. Liu, J., Liu, X., Yang, Y., Liao, Q. & Xia, Y. Contrastive Multi-View kernel learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **45** (8), 9552–9566. https://doi.org/10.1109/TPAMI.2023.3253211 (2023).
18. Zhao, N. et al. Kernel machine methods for integrative analysis of genome-wide methylation and genotyping studies. *Genet. Epidemiol.* **42** (2), 156–167. https://doi.org/10.1002/gepi.22100 (2018).
19. Bobb, J. F. et al. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics* **16** (3), 493–508. https://doi.org/10.1093/biostatistics/kxu058 (2015).
20. Zimmer, V. A., Lekadir, K., Hoogendoorn, C., Frangi, A. F. & Piella, G. A framework for optimal kernel-based manifold embedding of medical image data. *Comput. Med. Imaging Graph.* **41**, 93–107. https://doi.org/10.1016/j.compmedimag.2014.06.001 (2015).
21. Lu, C. K. & Shafto, P. Conditional deep Gaussian processes: Multi-Fidelity kernel learning. *Entropy (Basel)*. **23** (11), 1545. https://doi.org/10.3390/e23111545 (2021). Published 2021 Nov 20.
22. Krasnov, D. et al. A review of applications in breast Cancer detection. *Entropy (Basel)*. **25** (7), 1021. https://doi.org/10.3390/e25071021 (2023). Published 2023 Jul 4.
23. Latif, G., Alghazo, J., Sibai, F. N., Iskandar, D. N. F. A. & Khan, A. H. Recent advancements in fuzzy C-means based techniques for brain MRI segmentation. *Curr. Med. Imaging.* **17** (8), 917–930. https://doi.org/10.2174/1573405616666210104111218 (2021).
24. Wang, H. et al. ARM-Linux-Embedded system combined with magnetic resonance imaging for progression prediction of brain tumors. *Comput. Math. Methods Med.* **2022**, 4224749. https://doi.org/10.1155/2022/4224749 (2022). Published 2022 Mar 15.
25. Hong, Z. Q. & Yang, J. Y. Lung cancer. UCI machine learning repository. *Pattern Recogn.* **24** (4), 317–324. https://doi.org/10.24432/C57596 (1991).
26. National Lung Screening Trial Research Team et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New. Engl. J. Med.* **365**, 395–409. https://doi.org/10.1056/NEJMoa1102873 (2011).
27. Bakr, S. et al. Data for NSCLC Radiogenomics (Version 4) [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2017.7hs46erv (2017).
28. Li, Z., Wang, B., Li, J., Hua, Y. & Zhang, S. Local differential privacy protection for wearable device data. *PLoS One*. **17** (8), e0272766. https://doi.org/10.1371/journal.pone.0272766 (2022). Published 2022 Aug 17.
29. Bu, Z., Dong, J., Long, Q. & Su, W. J. Deep learning with Gaussian differential privacy. *Harv. Data Sci. Rev.* **2020** (23). https://doi.org/10.1162/99608f92.cfc5dd25 (2020).
30. Wang, T., Zhang, X., Feng, J. & Yang, X. A Comprehensive Survey on Local Differential Privacy toward Data Statistics and Analysis. *Sens. (Basel)* ;**20**(24):7030. (2020). Published 2020 Dec 8. doi:https://doi.org/10.3390/s20247030.
31. Almadhoun, N., Ayday, E. & Ulusoy, Ö. Differential privacy under dependent tuples-the case of genomic privacy. *Bioinformatics* **36** (6), 1696–1703. https://doi.org/10.1093/bioinformatics/btz837 (2020).
32. Singh, C., Ranade, S. K., Kaur, D. & Bala, A. An intuitionistic fuzzy C-Means and local Information-Based DCT filtering for fast brain MRI segmentation. *J. Imaging Inf. Med. Published Online April.* **22** https://doi.org/10.1007/s10278-023-00899-6 (2024).
33. Zhao, X., Chen, H., Li, B., Yang, Z. & Li, H. Using fuzzy C-Means clustering to determine first arrival of microseismic recordings. *Sens. (Basel)*. **24** (5), 1682. https://doi.org/10.3390/s24051682 (2024). Published 2024 Mar 5.
34. Tovar-Falón, R. & Martínez-Flórez, G. A new class of exponentiated Beta-Skew-Laplace distribution. *Acad. Bras. Cienc.* **94** (4), e20191597. https://doi.org/10.1590/0001-3765202220191597 (2022). Published 2022 Oct 24.
35. Soltanifar, M., Escobar, M., Dupuis, A., Chevrier, A. & Schachar, R. The asymmetric Laplace Gaussian (ALG) distribution as the descriptive model for the internal proactive Inhibition in the standard stop signal task. *Brain Sci.* **12** (6), 730. https://doi.org/10.3390/brainsci12060730 (2022). Published 2022 Jun 1.
36. Guo, P. et al. Radial basis function interpolation supplemented lattice Boltzmann method for electroosmotic flows in microchannel. *Electrophoresis* **42** (21–22), 2171–2181. https://doi.org/10.1002/elps.202100155 (2021).
37. Sousa Júnior, E., Freitas, A., Rabelo, R. & Santos, W. Estimation of Radial Basis Function Network Centers via Information Forces. Entropy (Basel). ;24(10):1347. Published 2022 Sep 23. (2022). https://doi.org/10.3390/e24101347
38. Shorten, R. & Murray-Smith, R. Side effects of normalising radial basis function networks. *Int. J. Neural Syst.* **7** (2), 167–179. https://doi.org/10.1142/s0129065796000130 (1996).
39. Derry, A., Krzywinski, M. & Altman, N. Convolutional neural networks. *Nat. Methods*. **20** (9), 1269–1270. https://doi.org/10.1038/s41592-023-01973-1 (2023).
40. Soffer, S. et al. Convolutional neural networks for radiologic images: A radiologist's guide. *Radiology* **290** (3), 590–606. https://doi.org/10.1148/radiol.2018180547 (2019).
41. Xu, Y. & Zhang, H. Convergence of deep convolutional neural networks. *Neural Netw.* **153**, 553–563. https://doi.org/10.1016/j.neunet.2022.06.031 (2022).
42. Larner, A. J. Accuracy of cognitive screening instruments reconsidered: overall, balanced or unbiased accuracy? *Neurodegener Dis. Manag.* **12** (2), 67–76. https://doi.org/10.2217/nmt-2021-0049 (2022).
43. Steinley, D., Brusco, M. J. & Hubert, L. The variance of the adjusted Rand index. *Psychol. Methods*. **21** (2), 261–272. https://doi.org/10.1037/met0000049 (2016).
44. Zhao, Y. et al. Local Differential Privacy based Federated Learning for Internet of Things. dio: (2020). https://doi.org/10.1109/JIOT.2020.3037194
45. Truex, S. et al. LDP-Fed: Federated learning with local differential privacy. dio: (2020). https://doi.org/10.1145/3378679.3394533
46. Yang, J. et al. Matrix Gaussian Mechanisms for Differentially-Private Learning. (2023). https://doi.org/10.1109/TMC.2021.3093316
47. Sun, L., Qian, J. & Chen, X. Ldp-fl: practical private aggregation in federated learning with local differential privacy.2020. https://doi.org/10.24963/ijcai.2021/217
48. Liu, R. et al. Fedsel: Federated sgd under local differential privacy with top-k dimension selection. Dio: (2020). https://doi.org/10.1007/978-3-030-59410-7_33
49. Patil, R. S., Kulkarni, S. B. & Gaikwad, V. L. Artificial intelligence in pharmaceutical regulatory affairs. *Drug Discov Today*. **28** (9), 103700. https://doi.org/10.1016/j.drudis.2023.103700 (2023).
50. Price, W. N. 2, Cohen, I. G. & nd, Privacy in the age of medical big data. *Nat. Med.* **25** (1), 37–43. https://doi.org/10.1038/s41591-018-0272-7 (2019).
51. Kroes, S. K., Janssen, M. P., Groenwold, R. H. & van Leeuwen, M. Evaluating privacy of individuals in medical data. *Health Inf. J.* **27** (2), 1460458220983398. https://doi.org/10.1177/1460458220983398 (2021).
52. Liu, R., Yang, J. & Wu, J. When big data backfires: the impact of a perceived privacy breach by pharmaceutical E-Retailers on customer boycott intention in China. *Int. J. Environ. Res. Public. Health*. **19** (8), 4831. https://doi.org/10.3390/ijerph19084831 (2022). Published 2022 Apr 15.
53. Avraam, D. et al. Privacy preserving data visualizations. *EPJ Data Sci.* **10** (1), 2. https://doi.org/10.1140/epjds/s13688-020-00257-4 (2021).
54. Shrivastava, U., Song, J., Han, B. T. & Dietzman, D. Do data security measures, privacy regulations, and communication standards impact the interoperability of patient health information? A cross-country investigation. *Int. J. Med. Inf.* **148**, 104401. https://doi.org/10.1016/j.ijmedinf.2021.104401 (2021).
55. Peng, S. et al. Blockchain data secure transmission method based on homomorphic encryption. *Comput. Intell. Neurosci.* https://doi.org/10.1155/2022/3406228 (2022). 2022:3406228. Published 2022 Apr 30.
56. Wang, W., Qin, T. & Wang, Y. Encryption-free data transmission and hand-over in two-tier body area networks. *Comput. Methods Programs Biomed.* **192**, 105411. https://doi.org/10.1016/j.cmpb.2020.105411 (2020).

57. Mohammad, B. A., Brennan, P. C. & Mello-Thoms, C. A review of lung cancer screening and the role of computer-aided detection. *Clin. Radiol.* **72**, 433–442. https://doi.org/10.1016/j.crad.2017.01.002 (2017).
58. Armato, S. G. 3 et al. Lung cancer: performance of automated lung nodule detection applied to cancers missed in a CT screening program. *Radiology* **225**, 685–692. https://doi.org/10.1148/radiol.2253011376 (2002).
59. Buty, M. et al. In: Medical image computing and computer-assisted intervention. Ourselin S., Joskowicz L., Sabuncu M., Unal G., Wells W., editors. Springer; Cham: Characterization of lung nodule malignancy using hybrid shape and appearance features. pp. 662–70. (2016).
60. Hussein, S., Cao, K., Song, Q. & Bagci, U. Risk stratification of lung nodules using 3D CNN-based multi-task learning. In *Information Processing in Medical Imaging* (eds Niethammer, M. et al.) 249–260 (Springer, 2017).
61. Khosravan, N. et al. A collaborative computer aided diagnosis (C-CAD) system with eye-tracking, sparse attentional model, and deep learning. *Med. Image Anal.* **51**, 101–115. https://doi.org/10.1016/j.media.2018.10.010 (2019).
62. Ciompi, F. et al. Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. *Med. Image Anal.* **26**, 195–202. https://doi.org/10.1016/j.media.2015.08.001 (2015).
63. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Commun. ACM*. **60**, 84–90 (2017).
64. Sermanet, P. et al. OverFeat: integrated recognition, localization and detection using convolutional networks. *ArXiv* (2014). 1312.6229.
65. Gu, X., Wang, Y., Chan, T. F., Thompson, P. M. & Yau, S. T. Genus zero surface conformal mapping and its application to brain surface mapping. *Inf. Process. Med. Imaging.* **18**, 172–184. https://doi.org/10.1007/978-3-540-45087-0_15 (2003).
66. Venkadesh, K. V. et al. Deep learning for malignancy risk Estimation of pulmonary nodules detected at low-dose screening CT. *Radiology* **300**, 438–447. https://doi.org/10.1148/radiol.2021204433 (2021).
67. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *IEEE Conf. Comput. Vis. Pattern Recognit.* :770–778. (2016).
68. Szegedy, C. et al. Going deeper with convolutions. *IEEE Conf. Comput. Vis. Pattern Recognit.* :1–9. (2015).
69. Ocampo, P. et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *J. Thorac. Oncol.* **13**, S562. https://doi.org/10.1038/s41591-018-0177-5 (2018).
70. Lin, C. H., Lin, C. J., Li, Y. C. & Wang, S. H. Using generative adversarial networks and parameter optimization of convolutional neural networks for lung tumor classification. *Appl. Sci.* **11**, 480 (2021).
71. Ren, Z., Zhang, Y. & Wang, S. A hybrid framework for lung cancer classification. *Electronics* **11**, 1614. https://doi.org/10.3390/electronics1010000 (2022).
72. Le, T. T. et al. Differential privacy-based evaporative cooling feature selection and classification with relief-F and random forests. *Bioinformatics* **33** (18), 2906–2913. https://doi.org/10.1093/bioinformatics/btx298 (2017).
73. Gupta, R., Saxena, D., Gupta, I., Makkar, A. & Kumar Singh, A. Quantum machine learning driven malicious user prediction for cloud network communications. in *IEEE Netw. Lett.*, **2022**:4: 174–178. https://doi.org/10.1109/LNET.2022.3200724
74. Gupta, R., Saxena, D., Gupta, I. & Singh, A. K. Differential and triphase adaptive Learning-Based Privacy-Preserving model for medical data in cloud environment. in *IEEE Netw. Lett.*, **2022**:4: 217–221. https://doi.org/10.1109/LNET.2022.3215248

Acknowledgements

The authors thank Professor Liu CY for help in data analysis. The authors thank Dr. Hao BB for suggestions and corrections that improved the text.

Author contributions

H.Y contributed to article writing. Z.M, Z.H, and W.N designed the study and guided the experiment. Y.M devoted themselves to data collection. K.M provided fund support. Z.M provides technical support.All authors were responsible for experimental design and proofread the final version of manuscript.

Funding

This work has been supported by the Nanjing Health Department Medical Technology Development Foundation (YKK20171) and Nanjing Health Department Medical Technology Development Foundation (YKK23249).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-01873-8.

Correspondence and requests for materials should be addressed to Z.M.