Citation:

Kuang, Yongxiang, Bin Jiang, Xuerong Cui, Shibao Li, Yongxin Liu, and Houbing Song. "Flexible Differential Privacy for Internet of Medical Things Based On Evolutionary Learning." IEEE Internet of Things Journal, 2024, 1–1. https://doi.org/10.1109/JIOT.2024.3366889.

DOI:

https://doi.org/10.1109/JIOT.2024.3366889

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3366889

Flexible Differential Privacy for Internet of Medical Things Based on Evolutionary Learning

Yongxiang Kuang, Bin Jiang, *Member, IEEE*, Xuerong Cui, *Member, IEEE*, Shibao Li, *Member, IEEE*, Yongxin Liu, *Senior Member, IEEE*, and Houbing Song, *Fellow, IEEE*

***Abstract*—With the development of Internet of Medical Things(IOMT), a lot of medical data are stored and released for both scientific research and practical applications. Accurate medical data is very valuable, but it also brings a huge risk of privacy leakage. Moreover, improving the privacy of data often leads to the reduction of data validity. Privacy and effectiveness are in conflict, and their balance is a typical multi-objective optimization problem (MOP). In this paper, we try to use differential privacy to disturb medical data to protect personal privacy. We propose the Environment Switching Algorithm (ESA) based on evolutionary learning to solve this MOP. ESA has excellent performance, which can ensure convergence speed and optimization performance at the same time. The result of optimization is a pareto front (PF) of huge scale, which includes solutions with different characteristics. We put forward a method of double clustering to select the appropriate solution from PF. Based on the above, we conclude the whole method as Flexible Differential Privacy Algorithm based on Evolutionary Learning (FDPEL). FDPEL can realize flexible differential privacy for medical data, while ensuring data privacy and data validity. FDPEL is suitable for privacy protection of medical data of different scales, which makes it have a practical applications value.**

***Index Terms*—Internet of Medical Things, Differential privacy, Multi objective optimization, Evolutionary learning, Pareto frontier.**

I. INTRODUCTION

WITH the rapid development of information technology, the Internet of Medical Things (IOMT) has been developed rapidly [1], as shown in Fig.1. IOMT can collect rich medical information by using various sensing, storage and communication modules, and use the information for online diagnosis and data analysis [2]. In the medical field, the existing medical data is of great value for scientific research and disease diagnosis [3]. For example, scholars can use machine learning (ML) and other technologies to analyze patients' basic information and daily life habits, and find out their relationship with a certain disease, thus helping others to screen the potential risk of diseases [4].

IOMT system is vulnerable to attack, and it will cause privacy leakage when the patient's personal information is identified, which will affect personal life [5]. Therefore, the privacy protection of IOMT has always been a challenging topic. The privacy leakage of IOMT can occur at all stages [6]. This paper focuses on the privacy protection of data publishing. Many medical institutions and official organizations will try to publish the data generated by IOMT on the Internet for scholars to conduct data analysis and scientific research. The index that can directly identify individuals (such as names) will be hidden when data is published, and K-anonymization and other methods will be tried to protect privacy [7]. However, there are also some malicious attacks, such as chain attacks, which can identify individual users and seriously damage people's privacy [8]. In particular, differential privacy has been widely used and become a privacy protection standard, because it has strict mathematical proof [9]. In this paper, differential privacy is used to protect the privacy of medical data in IOMT.

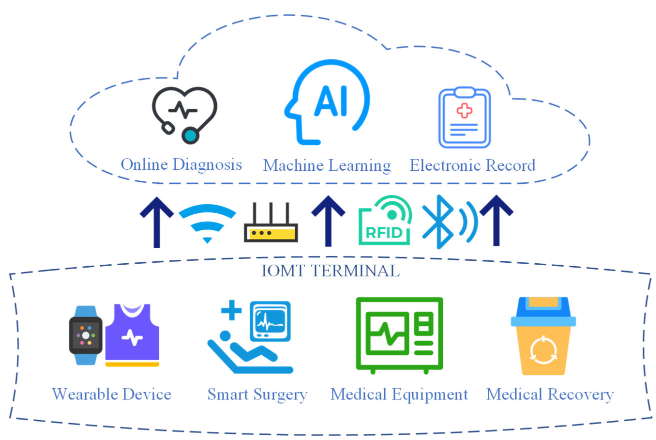


Fig. 1: Internet of Medical Things(IOMT)

In privacy protection, the privacy and validity of data often conflict [10]. Generally speaking, the higher the degree of privacy protection, the less valuable it is to scientific research, that is, the less effective it is. Similarly, the more valuable the data, the higher the risk of privacy leakage. Data publishers must balance data privacy and effectiveness [11]. The method of balance is often carried out according to subjective intention, and it will be ineffective when a large amount of data is published frequently.

This work was supported in part by National Natural Science Foundation of China under Grant 62102264, Young Expert Project of Taishan Scholar under Grant tsqnz20230602, Youth Innovation University Team Project in Shandong under Grant 2022KJ062, and Independent Innovation Fund of China University of Petroleum (East China) under Grant 22CX06056A. *(Corresponding author: Bin Jiang.)*

Yongxiang Kuang, Bin Jiang, Xuerong Cui and Shibao Li are with College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao 266580, China.(email:2004010608@s.upc.edu.cn; jiangbin@upc.edu.cn; cxr@upc.edu.cn; Lishibao@upc.edu.cn).

Yongxin Liu is with the Department of Mathematics, Embry-Riddle Aeronautical University, Daytona Beach, FL 32114 USA (e-mail: LIUY11@erau.edu).

Houbing Song is with the Department of Information Systems, University of Maryland, Baltimore County (UMBC), Baltimore, MD 21250 USA (email: h.song@ieee.org; songh@umbc.edu).

The balance between validity and privacy is a typical Multi-objective Optimization Problem (MOP), and evolutionary algorithm (EA) is often used to solve MOP [12]. EA can find a set of trade-off solutions, which are optimal when considering all conflicting objectives. However, the scale of health topic dataset is variable, and the number of decision variables has expanded from a few to thousands. Traditional evolutionary algorithms, such as Non-dominated Sorting Genetic Algorithm-II (NSGA-II) and Particle Swarm Optimization (PSO), have a sharp decline in performance and become very difficult to converge when dealing with decision variables exceeding 100 [13]. Therefore, they are not applicable in the context of this article. In order to accelerate the convergence, scholars try to combine EA with ML to improve its performance in large-scale multi-objective optimization, which is defined as evolutionary learning. However, due to its emphasis on convergence speed, when it reaches a certain degree of convergence, the general evolutionary learning algorithms are easy to fall into local optimum, which is also considered and Environment Switching Algorithm (ESA) is proposed in this paper.

Like other EA, the population size of evolutionary learning algorithms is usually set to a relatively large value, such as 100-1000. The difficulty of filtering caused by unclear user preferences needs to be solved. Considering it, a double clustering method is proposed.

To sum up, we propose an automatic privacy protection method, which we call Flexible Differential Privacy Algorithm based on Evolutionary Learning (FDPEL).

We model the privacy-preserving of medical data as a MOP and solve it with a multi-objective optimization algorithm based on evolutionary learning. The contributions of this paper are as follows:

1) To achieve excellent optimization on medical datasets of different scale, we designed Environment Switching Algorithm (ESA) based on evolutionary learning. ESA can solve MOP well when facing medical datasets, so as to achieve flexible differential privacy purpose. ESA focuses on both convergence and optimization performance.

2) To evaluate the effect of differential privacy directly, we put forward an evaluation system for disturbed medical dataset. After using differential privacy to disturb the original data, a series of objective functions are designed to evaluate the privacy and validity of the disturbed data.

3) In order to help users filter out appropriate optimization results and realize flexible differential privacy on medical dataset, we propose a double clustering method. This method is convenient for users to select the suitable solution vectors from the large-scale solution sets. The solution set is divided from two perspectives: one is the PF performance generated from ESA, and the other is the privacy budget set of indexes.

The rest of this paper is organized as follows: In Section.II, we introduce the related works. In Section.III, we describe FDPEL in detail. In Section.IV, we evaluate the effectiveness and universality of FDPEL through experiments. Finally, we give a conclusion.

II. Background and Related Work

In this section, we first discuss and analyze the model and application of differential privacy. Then, we summarize and discuss the multi-objective optimization problem and evolutionary learning. Finally, we investigate the selection of optimization results in practical problems.

A. Differential Privacy for Medical Data

Differential privacy is a privacy protection model with strict theoretical proof. Assuming D is the dataset to be released, it contains both numerical and non-numerical columns, and we are attempting to add noise to it.

Assuming ε is a positive real number, and A is a random algorithm that takes a dataset as input. S represents all the outputs of algorithm A on dataset D and $D^{'}$. If the random algorithm A satisfies ε-differential privacy, then

Definition1 Differential Privacy

$$\mathcal{P}\left[A\left(D \in S\right)\right] \leq e^{\varepsilon} \times \mathcal{P}\left[A\left(D^{'} \in S\right)\right] \tag{1}$$

ε represents the privacy budget, and the smaller ε is, the higher the degree of privacy protection [14].

Differential privacy has been widely used in the protection of various data, and achieved remarkable results [15]. In [16], Jiang $et\ al.$ propose a new federated edge learning framework based on hybrid differential privacy and adaptive compression for industrial data processing. In [17], Jiang $et\ al.$ discuss how differential privacy is applied to social network analysis, and analyzes privacy attacks and differential privacy models in social networks. A trie-based iterative statistic method, which combines additive secret sharing and local differential privacy technologies, was proposed in [18] to protect real-time location information.

Many scholars try to use various methods to protect the privacy of each module of IOMT [19]. In [20], a blockchain-based two-stage federated learning approach is proposed, which allows IOMT devices to cooperatively train the global model without collecting data to the central server, thus reducing the risk of privacy leakage. In [21], Jia $et\ al.$ present two privacy-preserving authentication protocols for IOMT based on elliptic curve cryptography (ECC) and physically unclonable functions (PUFs), respectively, in terms of the capacity of involved entities. In [22], Zeng $et\ al.$ propose an efficient partially-policy-hidden and large universe ABE scheme with public traceability to construct a practical IOMT system. Differential privacy has been widely used to protect medical data since it was put forward. In [23], Gupta $et\ al.$ propose a novel Differential and TriPhase adaptive learning-based Privacy-Preserving Model (DT-PPM) for medical data protection by enabling secure data storage, analysis, and sharing in the cloud environment. In [8], Wang $et\ al.$ propose a privacy-enhanced disease diagnosis mechanism using federated learning(FL) based on differential privacy for the IOMT.

B. Multi-Objective Optimization and Evolutionary Learning

Without loss of generality, a Multi-objective Optimization Problem(MOP) without constraints can be modeled as:

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3366889

$$Minimize\ F(x) \in Y, x \in \alpha \tag{2}$$

where $F(x) = (f_1(X), f_2(X), \ldots, f_m(X))$ and and $X = (x_1, x_2, \ldots, x_n)$. Respectively, m represents the number of the objective function in the objective space Y, and n represents the number of the decision variable X in the search space α. Large-Scale Multi-objective Optimization Problem(LMOP) is MOP when $n \geq 1000$ and $m \geq 2$. Some scholars have applied the multi-objective optimization method to medical problems. Most optimization problems in the real world are MOP or LMOP [24]. In [25], Zhou $et\ al.$ propose a multi-objective based feature selection (MO-FS) algorithm for Lesion Malignancy Classification.

In the past decades, many classic multi-objective evolutionary algorithms(MOEAs) such as NSGA-II and algorithms based on them have been proposed and widely used to solve practical problems [26]. For example, in [27], Ding $et\ al.$ use MOEAs to addresses a flexible job shop scheduling problem under time-of-use electricity tariffs with the objective of minimizing total energy consumption while considering a predefined makespan constraint.

When dealing with LMOP, the effect of conventional MOEAs will drop sharply [28]. In order to solve this kind of problem, many advanced algorithms have been put forward. These algorithms can be roughly divided into two categories [29]. The first is to optimize decision variables by using various methods such as grouping and clustering. For example, in [30], Xu $et\ al.$ propose a new metric called the optimization degree of the convergence-related decision variable to each objective to calculate the contribution objective of each decision variable. The second is to introduce the idea of machine learning(ML) [31], which we call evolutionary learning.

However the generality of the algorithm is required in this scenario. The number of decision variables in the data set of health topics may be several or hundreds. Using conventional MOEAs to solve it will lead to difficulty in convergence. Using ML-based evolutionary learning algorithm will lead to over-emphasis on accelerating convergence and decrease the diversity of solutions. This is unacceptable to us. Therefore, in this paper, we propose a more flexible algorithm based on evolutionary learning.

C. Selection of Optimization Results

In the solution of MOPs, the population number is usually set to 100 or more. The solution is not a function value, but a target vector. Several objective functions considered at the same time are often in conflict, and optimizing one objective function alone will make other objective functions worse. Therefore, the two solutions of multi-objective optimization are often not directly comparable. A solution performs well on one objective function, but poorly on other objective functions. Therefore, we often use dominance relation to compare two solutions.

Given the target vector $F = (f_1, f_2, \cdots, f_m) : \mathcal{X} \rightarrow \mathbb{R}^m$, where $\mathcal{X}$ is feasible solution space, $\mathbb{R}^m$ is target vector space, for solution x and $x^{'} \in \mathcal{X}$, if $f_i(x) \geq f_i(x^{'})$ for any $1 \leq i \leq m$, then x dominance $x^{'}$.

Based on the dominance relation, the result of multi-objective optimization is no longer unique, but a set of Pareto optimal solutions. For a solution x, if there is no other solution dominating x in $\mathcal{X}$, then x is called Pareto optimal. The set of objective vectors of all Pareto optimal solutions is called Pareto Front(PF). PF can help us to select the solutions preliminarily. But in practice, a solution vector is used. How to select the appropriate solution vector from the huge solution set? This requires additional processing of the solution set.

In [30], Zhang $et\ al.$ use clustering method to process the optimization results, but only divided the results into two categories. In [32], Hua $et\ al.$ choose a balance point in PF as the best point for analysis according to the preference of practical problems, this method is suitable for specific data sets and problems, but not universal. In [33], Xie $et\ al.$ use fuzzy decision method to select the optimization results, but the setting of weights often plays a decisive role, so it is not universal. In [34], Xu $et\ al.$ choose the point at the inflection point of PF as the optimal solution. Most papers only evaluate the PF curve, and then choose the optimization result according to the weight of preference. We believe that the solution users are seeking should be diverse and comprehensive, so we think it is not universally applicable.

III. Flexible Differential Privacy Algorithm based on Evolutionary Learning

In this section, we describe the details of Flexible Differential Privacy Algorithm based on Evolutionary Learning(FDPEL). FDPEL provides privacy-preserving for the publish of medical data. The goal of publishing dataset on health topics is to find a scientific way to accurately predict the probability of getting sick. These dataset are typically made up of n attribute columns such as some physical characteristics, daily habits, and one decision column such as whether you have a certain disease. FDPEL serves two primary purposes: 1) Safeguarding data privacy and prevent unauthorized identification of individuals through methods such as linkage attacks. 2) Maintain the validity of data and prevent it from becoming invalid due to disturbance because of differential privacy, resulting in wrong scientific research results. This section introduces our work from four aspects. Firstly, we describe the differential privacy process for medical data. Secondly, we design an evaluation system for differential privacy effect. Thirdly, we propose ESA, an algorithm based on evolutionary learning to improve the flexibility of differential privacy. Finally, we design a double clustering method for users to filter the appropriate optimization results.

A. Differential Privacy Process

Differential privacy is defined as Eq.1, where ε represents the privacy budget, and the smaller ε is, the higher the degree of privacy protection. The noise mechanism is the primary technique for achieving differential privacy protection, with commonly used noise addition mechanisms being the Laplace mechanism and the exponential mechanism. The amount of noise required for algorithms based on different noise mechanisms and satisfying differential privacy is closely related to the Global Sensitivity.

Definition2 Global Sensitive

For any function $f : D \to R^d$, The global sensitivity of function f is defined as

$$\triangle f = \max_{D,D'} \left\| f(D) - f(D') \right\|_1 \tag{3}$$

where D and D' differ by at most one record, R represents the mapped real number space, d represents the query dimension of the function f, and p represents the L_1 distance used to measure $\triangle f$.

In this paper, we mainly use Laplace mechanism to add noise, and the noise generated by Laplace distribution disturbs the real value to realize differential privacy protection. Consider the laplace distribution with a mean value of 0 and a scale parameter of b as $lap(x \mid b) = \frac{1}{2b} e^{\left(-\frac{|x|}{b}\right)}$. When the scale parameter $b = \frac{f}{\varepsilon}$, ε-differential privacy can be satisfied. ε is called the privacy budget. When $\triangle f$ is constant , the larger the privacy budget, the smaller the scale parameter b and the smaller the added noise. It can be found that the privacy budget ε is a very sensitive value to the noise disturbance level. In this work, we adjust the value of privacy budget of each attribute column to achieve the different degree of noise disturbance for different attribute.

In order to evaluate the noise-adding effect uniformly, it is necessary to standardize each column first. The specific implementation method is: for a certain column of data $X = \{x_1, x_1 \cdots x_n\}$, standardize X to get X_{norm}, $X_{norm} = \{x_{1norm}, x_{2norm} \cdots x_{nnorm}\}$, where

$$X_{inorm} = norm \times \frac{x_i - x_{imin}}{x_{imax} - x_{imin}} \tag{4}$$

next, add noise to X_{norm} to get X_{noisy}.

$$\begin{aligned} X_{\text{noisy}} = \Bigg\{ & x_{1\text{norm}} + \text{lap}\left(\frac{\Delta f}{\varepsilon_1}\right), x_{2\text{norm}} + \text{lap}\left(\frac{\Delta f}{\varepsilon_2}\right), \\ & \ldots, x_{n\text{norm}} + \text{lap}\left(\frac{\Delta f}{\varepsilon_n}\right) \Bigg\} \end{aligned} \tag{5}$$

The published data is X_{noisy}, and then the X_{noisy} is denormalized to get X_{pub}, which is the final published data.

$$\begin{aligned} X_{\text{noisy}} = \Big\{ & x_{1\text{pub}}, x_{2\text{pub}}, \ldots, x_{n\text{pub}} \mid \\ & x_{i\text{pub}} = x_{i\text{norm}}(x_{i\text{max}} - x_{i\text{min}}) + x_{i\text{min}} \Big\} \end{aligned} \tag{6}$$

B. Evaluation System of Noise Adding Effect

After differential privacy noise, we try to design a series of objective functions to evaluate the effect of differential privacy noise, as shown in Fig.2. As mentioned at the beginning of this chapter, we have two purposes: 1) to protect data privacy and prevent intruders from identifying the real recorded person according to vicious methods. 2) to maintain the validity of data and prevent it from becoming invalid due to privacy noise, resulting in wrong scientific research results. These two purposes are designed as two objective functions. The first one is called privacy function, which describes the degree of privacy protection of published data. The second one, which we call the validity function, describes the effectiveness of scientific research in publishing data. It should be noted that the smaller the values of the two objective functions, the higher the performance. The specific function design is as follows.

1) : Regarding the privacy evaluation of published data, we focus on the ε parameter of Laplace mechanism privacy budget and the evaluation of information retention after disturbance. These parameters need to directly reflect the degree of noise disturbance. In addition, the lower the information retention, and the higher the privacy of users is protected. This function is mainly composed of three items, Parameter of ε, Individual Item Retention and Overall Information Retention.

a) Parameter ε: The probability density function of laplacian noise added in this paper is $lap(x \mid b) = \frac{1}{2b} e^{\left(-\frac{|x|}{b}\right)}$. From the above analysis, it can be seen that the privacy budget ε is a sensitive value, and the degree of noise disturbance can be adjusted by ε. In this paper, each column has a separate privacy budget. By adjusting ε value, we can add noise to different attributes in different degrees.

For an attribute, the smaller the ε, the greater the disturbance to the original data, and the privacy of users is well protected. Therefore, we set the first item of the first function as

$$f_{11}(X, X_{pub}) = \sum_{i=1}^{n} \varepsilon(i) \tag{7}$$

where $\varepsilon(i)$ represent the privacy budget for the i-th attribute column. Because the data of each column has been standardized before adding noise, the weight of each attribute is same, and $f_{11}(X, X_{pub})$ can well represent the total disturbance.

b) Column Information Retention: The second parameter is set to Information Retention (single column), which is to take out each column separately and evaluate the total amount of information retained. In other words, when the information in the same column has not changed, the most information is retained, and the value of this function reaches the maximum. When the information changes greatly, the value of this function is 0. When the value of this parameter becomes smaller, it indicates the improvement of privacy.

For Categorical and Integer columns, we define the single-column information retention as

$$IIR_{\text{C}} = \frac{\sum_{i \in A} \left(\|X_i\| \max(X_i) - \sum_{j=1}^{\|X_i\|} |X_i(j) - x_{\text{ipub}}(j)| \right)}{k_1 \max(X_i)} \tag{8}$$

$$IIR_{\text{I}} = \frac{\sum_{i \in B} \left(\sum_{j=1}^{||X_i||} |X_i(j) - X_{ipub}(j)| \leq \rho \right)}{k_2} \tag{9}$$

where A represents the set of Categorical column numbers and B represents the set of Integer column numbers. k_1 and k_2 respectively represent two adjustment parameters, adjusting the proportion of IIR_{C} and IIR_{I} in the evaluation.

To sum up, we get the final value of this parameter

$$f_{12}(X, X_{pub}) = IIR_{\text{C}} + IIR_{\text{I}} \tag{10}$$

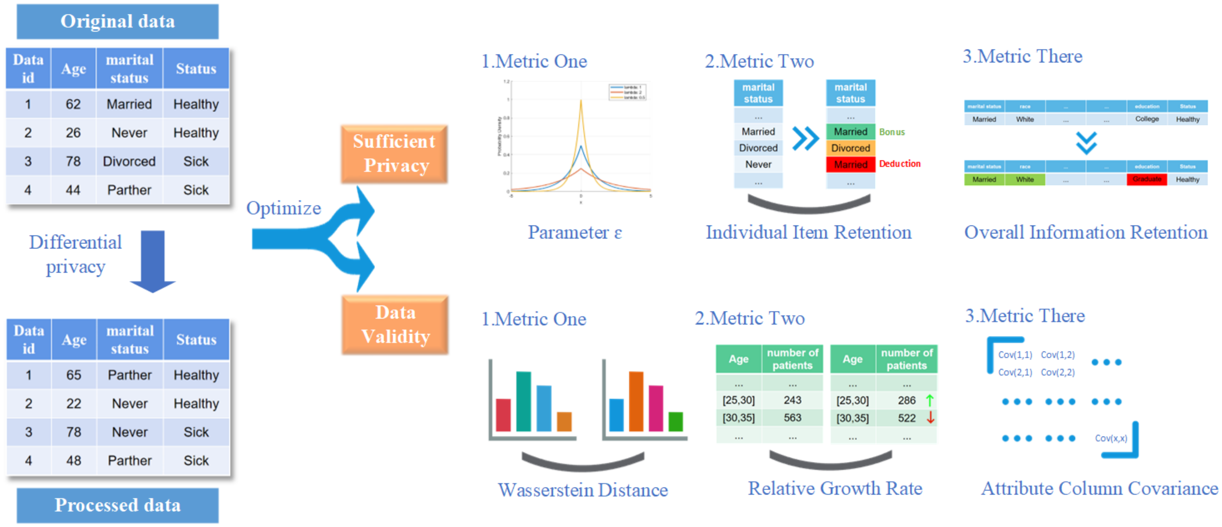


Fig. 2: Evaluation System of Noise Adding Effect

c) Overall Information Retention: The third parameter is to evaluate the total amount of information retained by each row, namely information retention of one data record. In practice, each data record represents all the information of a person, so we evaluate the degree of change of rows. For Categorical columns, there is no change after adding noise, so we count them. For Integer columns, the change after adding noise is less than a certain threshold, so we count them. When the result of counting is greater than another threshold, the total count is increased by one. To obtain the parameter value, add up all the counts and divide the sum by a weight parameter. Formulated as follows

$$f_{13}(X, X_{\text{pub}}) = \sum_{j=1}^{|X_1|} \sum_{i=1}^{n} [(i \in A \&\& X_i(j)) == X_{ipub}(j) \\ || \; ((i \in B \&\& X_i(j)) - X_{ipub}(j) \leq \rho) \geq \sigma] \tag{11}$$

where A represents the set of Categorical column numbers and B represents the set of Integer column numbers. X_1 represents a representative column. ρ represents the threshold when Integer column has not changed. σ represents the threshold of information retention for the whole line.

Finally, the privacy function is expressed as the sum of three parameters:

$$F_{\text{privacy}}(X, X_{\text{pub}}) = f_{11} + f_{12} + f_{13} \tag{12}$$

where f_{11}, f_{12}, f_{13} respectively represent parameter ε, column information retention and overall information retention.

2) Evaluation of Efficiency: Regarding the evaluation of data validity, we focus on some probability functions and the correlation of columns. In addition, data users often focus on the correlation between different attribute, the correlation between attributes and judgment results. Therefore, our function design is also based on this. The validity function is mainly divided into three terms, Wasserstein Distance, Relative Growth Rate and Attribute Covariance.

a) Wasserstein Distance: The first parameter of validity is Wasserstein Distance. This is a probability statistical method commonly used in the field of machine learning at present. It describes the minimum cost required to transform from one distribution to another. In our study, the smaller the Wasserstein Distance is, the closer the noise data is to the original data, which is more beneficial to scientific research. For P and Q distributions, the Wasserstein Distance is expressed as

$$W(P, Q) = \inf_{\gamma \in \prod(P,Q)} E(||x - y||) \tag{13}$$

where inf represents the largest lower boundary, and $\gamma \in \prod(P, Q)$ represents the joint distribution of P and Q. The value of this parameter is as follows.

$$f_{21}(X, X_{\text{pub}}) = \sum_{i=1}^{n} W(X_i, X_{ipub}) \tag{14}$$

where X_i, X_{ipub} respectively represent the i-th attribute of X, X_{pub}.

b) Relative Growth Rate: The second parameter of validity is Relative Growth Rate. In this parameter, we try to treat each attribute as a unit and cross-count it with the judgment column. We count the number of each value in the judgment column corresponding to the attribute value. The smaller the value of this number, the smaller the change of the number of judgment values corresponding to each attribute value, and the higher the effectiveness. For columns of Categorical type, cross statistics can be performed directly. For integer column, some mapping changes are needed first, and then cross statistics are carried out. We take a mapping interval $Map = [0 : \tau : upper]$, where τ represents the interval of mapping, and the smaller the value of τ, the more sensitive the result is to the change of Integer column. Next, we map X to $X_{\text{m}} = \{X_1 \rightarrow X_{1\text{m}}, X_2 \rightarrow X_{2\text{m}}, \ldots, X_n \rightarrow X_{n\text{m}}\}$. Finally, we carry out the following cross statistics to obtain the C_{cross} matrix, that is, the result of cross statistics.

$$C_{\text{cross}} = \begin{pmatrix} ||X_{\text{im}} = \mathrm{M}(1)\&X_{\text{p}} = 0|| & \cdots & ||X_{\text{im}} = \mathrm{M}(1)\&X_{\text{pre}} = m|| \\ \ldots & \ddots & \ldots \\ ||X_{\text{im}} = \mathrm{M}(n)\&X_{\text{p}} = 0|| & \cdots & ||X_{\text{im}} = \mathrm{M}(n)\&X_{\text{p}} = m|| \end{pmatrix} \tag{15}$$

where X_{p} represents the decision column. Finally, we subtract the values of two cross-statistical matrices to get $C = C_{cross} - C_{pubcorss}$, and then find the Frobenius norm of C to get the value of this parameter.

$$f_{22}(X, X_{\text{pub}}) = \sqrt{\sum_i \sum_j |c_{ij}|^2} \tag{16}$$

c) Attribute Covariance: The third parameter of validity is Attribute Column Covariance. In this parameter, we try to analyze the correlation between the attribute column and the judjement column. The covariance of two distributions can well describe the correlation between two variables. Find the covariance matrix of the original data and the data after noise, and get C_{cov} and C_{pubcov}. After subtracting the corresponding elements to get c_{ij}, find the Frobenius norm of the result C, that is the value of this parameter.

$$C_{\text{cov}} = \begin{pmatrix} Cov(X_1, X_1) & \cdots & Cov(X_1, X_n) \\ \ldots & \ddots & \ldots \\ Cov(X_n, X_1) & \cdots & Cov(X_n, X_n) \end{pmatrix} \tag{17}$$

$$C_{\text{pubcov}} = \begin{pmatrix} Cov(X_{1pub}, X_{1pub}) & \cdots & Cov(X_{1pub}, X_{npub}) \\ \ldots & \ddots & \ldots \\ Cov(X_{npub}, X_{1pub}) & \cdots & Cov(X_{npub}, X_{npub}) \end{pmatrix} \tag{18}$$

$$f_{23}(X, X_{\text{pub}}) = \sqrt{\sum_i \sum_j |c_{ij}|^2} \tag{19}$$

Finally, the privacy function is expressed as the sum of three parameters:

$$F_{efficiency}(X, X_{\text{pub}}) = f_{21} + f_{22} + f_{23} \tag{20}$$

where f_{21}, f_{22} and f_{23} respectively represent wasserstein distance, relative growth rate and attribute covariance.

C. Environment Switching Algorithm Based on Evolutionary Learning

Without loss of generality, a Multi-objective Optimization Problem(MOP) without constraints can be modeled as Equ.2. In this question, m is 2, which means two objective functions. X contains a privacy budget of n attribute columns. Because the number of attributes in the dataset is different, n in the MOP is different. In some large dataset, when n is large, the conventional multi-objective optimization algorithm is difficult to converge quickly. However, some multi-objective optimization algorithms with fast convergence have poor diversity and it is difficult to achieve the best optimization effect. In order to solve this optimization problem, we propose the Environment Switching Algorithm Based on Evolutionary Learning (ESA), as Algorithm.1.

ESA is based on the ALMOEA [31] evolutionary learning framework, which is improved in this paper to make it more suitable for this MOP. We will introduce ESA according to the basic operating procedures. ESA is mainly divided into two stages, namely accelerating convergence stage and environmental switching and expanding diversity stage.

Algorithm 1 ESA

```
Input: m, n, FE_max
Output: the final population P
Initialize P, MLP;
while FE ≤ FE_max do
    Judge whether the evolution of P is stagnant;
    if P is evolving rapidly. then
        MLP ⇐ Training(MLP, P);
        Q ⇐ Reproduction(MLP, P);
        P ⇐ environment selection(P, Q);
    else
        (Stagnation for successive generations)
        Q ⇐ Reproduction(P); (Expand crossover and mutation)
        P ⇐ environment selection(P, Q); (Keep the promising P)
    end if
    FE = FE + N;
end while
return P;
```

1) Accelerated convergence stage: ESA is accelerated based on ALMOEA framework. In the initial stage, the population P and a multi layer perceptron(MLP) are generated by initialization. MLP is a strategy driven by Feedforward Artificial Neural Network to speed up the search in large-scale solution space. Feedforward neural network is a basic neural network structure, and the input of each layer directly comes from the output of the previous layer. The parameter optimization of MLP is updated by training the backward propagation of the gradient descent of the previous generation population P. When each generation of population P is generated, the algorithm will divide the population into Poor individuals and Elite individuals according to the evaluation of the objective function. Among them, Poor individual is the input of BP neural network, while Elite individual is the output of neural network. The training of MLP can be used to obtain the GDV of the input population P, so as to guide the population P to accelerate the convergence in the iteration. In the traditional evolutionary algorithm, the offspring population usually comes from the cross-recombination of the parents. This does not represent the fastest convergence direction of the population, so the convergence is slow. In the framework of ALMOEA, the new solution x^{new} can be obtained in the following ways:

$$x^{new} = x + r_1(x - x^{gdv}) + r_2(x^{d_1} - x^{d_2}) \tag{21}$$

where x^{gdv} is the learned GDV of x, which can be computed by inputting x into the trained MLP. Besides, d_1 and d_2 are two randomly selected solutions from the current population P. r_1 and r_2 are both random numbers ranging from 0 to 1. We apply the ALMOEA framework to the algorithm acceleration part of the evolutionary learning algorithm in this paper, so the r_1 value is usually set to be relatively large, and the solution set X can be accelerated by the learned GDV.

2) Environmental switching and expanding diversity stage: When the population develops to a certain stage, it will inevitably fall into convergence. At this time, it is difficult to distinguish the offspring population into Poor individuals and Elite individuals in the MLP training process under the framework of ALMOEA, and the original way of obtaining new solutions through GDV is also unsatisfactory. Therefore, it is necessary to judge whether the population has fallen into convergence at any time and reasonably change the way of generating offspring population. We judge the convergence by evaluating the inverted generative distance (IGD) of two Pareto surfaces.

$$IGD(X_{pre}, X_{new}) = \frac{1}{|X_{pre}|}\sqrt{\sum_{i=1}^{|X_{pre}|}(d_i^2)} \tag{22}$$

Among them, X_{pre} represents the previous generation population, and X_{new} represents the new population. The smaller the value of IGD, the higher the similarity between the two populations. When the new population is exactly the same as the original population, the value of IGD is 0. We can set a threshold value k, and the value of k can be determined according to the set number of populations produced in each generation. When the value of IGD is less than k, it can be judged that the new population is convergent at this time.

In order to avoid misjudgment, we set a counter count in the algorithm to record the number of generations that meet the condition $IGD \le k$. When the condition $IGD \le k$ is met, $count = count + 1$ is executed. When the condition is not met, execute $count = max\{count - 1, 0\}$. When the value of count reaches a certain threshold $k^{'}$, it can be judged that the population has reached the convergence state.

At this point, the function of the acceleration part of the algorithm is completed, and we try to switch the way of generating offspring population of the algorithm to expand population diversity. The method we adopt is to simulate binary crossover and polynomial mutation. By modulating the ratio of crossover and mutation, the diversity of solutions in future generations is improved.

The description of simulate binary crossover is as follows. Let P1 and P2 be two parent individuals, and C1 and C2 be crossed offspring individuals. We use binary codes to represent P1, P2, C1 and C2 respectively, and define $\beta = \frac{|C_1 - C_2|}{|P_1 - P_2|}$, which represents the ratio of direct distance between children and parents. Then the offspring can be represented as $C_1 = \frac{1}{2}(P_1 + P_2) - \frac{1}{2}(P_2 - P_1)$, $C_1 = \frac{1}{2}(P_1 + P_2) + \frac{1}{2}(P_2 - P_1)$.

It can be seen that β is a sensitive value for the generation of offspring, and the generation of β is derived from a probability distribution. When $\beta < 1$, the probability density $c(\beta) = \frac{1}{2}(n+1)\beta^n$, When $\beta > 1$, $c(\beta) = \frac{1}{2}(n+1)\beta^{n+2}$. The distribution function is $u = \int_0^{\overline{\beta}} c(\beta)\, d\beta$. Then

$$\beta = \begin{cases} (2u)^{\frac{1}{n+1}}, & if\ u \le 0.5 \\ (\frac{1}{2-2u})^{\frac{1}{n+1}}, & if\ u > 0.5 \end{cases} \tag{23}$$

The larger n is, the closer C1 and C2 are to P1 and P2. Therefore, by setting a smaller value of n, we can produce diverse solutions. However, when the value of n is set too small, the quality of the solution generated by the offspring will also decrease. Therefore, the value of n needs to be set in a reasonable range. Polynomial variation is described as follows. New solution $X_{\text{new}} = X_{\text{pre}} + \delta \cdot \Delta_{\text{max}}$.

$$\delta = \begin{cases} [2u + (1-2u)(1-\delta_1)^{\eta_m+1}]^{\frac{1}{\eta_m+1}-1}, & u \le 0.5 \\ 1 - [2 - 2u + 2(u-0.5)(1-\delta_2)^{\eta_m+1}]^{\frac{1}{\eta_m+1}}, & u > 0.5 \end{cases} \tag{24}$$

where $\delta_1 = (v_k - l_k)/(u_k - l_k)$, $\delta_2 = (u_k - v_k)/(u_k - l_k)$, u is a random number in an interval $[0, 1]$, and η_m is a distribution index selected by the user.

D. Double clustering evaluation method to select the results.

In general multi-objective work, the algorithm generates PF, means the optimization results is produced, where users can take one point from PF as final result according to their preferences. However, in the problems raised in this paper, it is difficult to divide privacy and effectiveness by a certain proportion. Therefore, we propose a method of double clustering, which is convenient for users to screen and produce the final desired results. There are two ranges: one is the PF performance generated by ESA, and the other is the set of privacy budget for each attribute.

1) Cluster based on PF performance: In this paper, there are two objective functions, which evaluate the privacy and validity of the noisy data set respectively, and finally generate PF. But on the same PF, the performance of individuals is also very different. Therefore, we try to divide individuals into k clusters according to their performance on two functions. For example, when k is set to 5, the original individuals can be classified into five categories: privacy-first, efficiency-first, privacy-focused, efficiency-focused and balance. The specific classification methods are as follows:

Among them, the distance is calculated by Euclidean distance, that is, $d(x, \mu) = \sqrt{\sum_{i=1}^{n}(x_i - \mu_i)^2}$, then the sum of squares of the distances from all sample points to the center of mass in a cluster is $TI = \sum_{j=0}^{m}\sum_{i=1}^{n}(x_i - \mu_i)^2$. The smaller the value of TI, the more similar the individuals in each cluster are, the better the clustering effect is. In the process of loop iteration, the value of TI is always getting smaller. This is actually an optimization problem. The value of k can be determined according to the division standard, and the larger the value of k, the finer the division. But when k is set too large, the purpose of clustering-to classify populations more clearly and clearly, will not be reflected. Generally, we suggest that the value of k be set to 3-7.

2) Cluster based on privacy budget set: According to the above work, we can know that each solution set X contains M solutions, where M is the population number of each generation set when running ESA. Among them, the i-th solution M_i is a collection of privacy budgets, $M_i = \{\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n\}$, where n is the number of attributes. We consider that users may have a preference for the noise level of a column when using the noisy dataset. In dataset with similar performance of privacy and efficiency, there may be different privacy budget preferences. For example, when the data user is sensitive to age information, they will tend to choose the solution with lower privacy budget of the age column to strengthen the protection of age while the overall privacy is the same.

The number of attributes in the privacy dataset is uncertain, which makes the cluster based on privacy budget maybe a high-dimensional clustering. If we continue to use the above clustering method, there are three shortcomings: (1) It is impossible to predict the number of clusters in advance, and when the value of k is set unreasonably, the reference value of the result is very small. (2) Clusters with arbitrary shapes can be identified, not just circles. (3) Some solutions with no common law can be identified as noise points. Therefore, we try to use a new density-based clustering method.

All the data in the dataset can be divided into three categories according to the density: core points, boundary points and noise points. They are distinguished by two parameters: the clustering radius E and the minimum number of points $MinPts$. The core store means that there are more than points in the radius E. A boundary point represents a point that falls within the Eps neighborhood of the core point, but is not the core point. Other points are as noise points.

The pseudo-code of the clustering algorithm is as Algorithm.2. In the loop iteration, all the uncollections are clustered. The number of clusters is not set in advance, but determined according to the density of solution distribution. By setting the parameters Eps and $MinPts$ reasonably, all solutions can be well classified according to the privacy budget.

3) Application of double clustering: The double clustering methods described above classify all the solutions by different methods and standards. The first one is clustering according to the value of the objective function, which reflects the privacy and effectiveness of the solution. The second is clustering according to the value of the solution, which reflects different privacy budget preferences. The remaining work is to combine the results of the two clustering methods and present them to users of FDPAEL. For example, in the first cluster, when the value of k is set to 5, it is divided into 5 categories, privacy-first, efficiency-first, privacy-focused, efficiency-focused and balance. In the second clustering, according to the different distribution of privacy budget, the algorithm divides the results into $n1 - n8$ and some single points.

As shown in Table.I, users can observe the result characteristics of the first cluster and the second cluster, find the corresponding population serial number, and add noise to their own datasets. Users can choose the solution they need from a huge population according to their own needs. This not only ensures the balance between effectiveness and privacy, but also ensures different privacy budget preferences.

Algorithm 2 Cluster based on privacy budget set

Input: solutions X, Eps, Minimum clustering points M
Output: A collection of clusters P
Initialize;
Mark all objects in X as unvisited.;
for Each object p in X **do**
  if p has been classified into a cluster or marked as noise. **then**
    Continue;
  else
    Check the Eps neighborhood $Nrp(p)$ of p;
    if $Nrp(p) \leq M$ **then**
      Marks p as a boundary point or a noise point;
    else
      Mark p as the core point, establish a new cluster C, and add all points in $Nrp(p)$ to C.
      for All unvisited objects in $Nrp(p)$ **do**
        Check its Eps neighborhood $Nrp(p)$, and if $Nrp(p)$ contains at least $MinPts$ objects, add the objects $Nrp(p)$ that do not belong to any cluster;
      end for
    end if
  end if
end for
return P;

TABLE I: Application of double clustering

ID	First clustering	Second clustering
1	privacy-focused	n_1
2	privacy-first	n_5
3	balanced	single point
4	efficiency-focused	n_8
...	...	...

IV. Experimental Results and Analysis

In this paper, FDPEL conducts flexible differential privacy on data, aiming at achieving the purpose of privacy protection while maintaining data validity.

In this section, we carried out experiments to verify the effectiveness of this method. This section starts from the following aspects: First, we run FDPEL through the whole process and show all the data in it. Secondly, we analyse the performance of FDPEL. Thirdly, we design a comparative experiment for multi-objective optimization algorithm ESA, and show the superiority of the ESA. Finally, we extend FDPEL to datasets of other medical topics to prove its applicability.

A. Presentation of Applied Datasets

There are three datasets applied in Experiment. The first one is used for Subsection.IV-B, and the other two are used for Subsection.IV-E.

The first dataset comes from the Mother's Significant Feature (MSF) dataset in IEEEDataPort [35]. MSF contains

450 records with a total of 130 attributes, including mother characteristics, father characteristics and health outcomes. The detailed dataset is created to understand the characteristics of mothers in three stages of reproductive age (adolescence, marriage and pregnancy). The dataset covers all possible complications related to children's health, mother's health and pregnancy outcome.

The second dataset comes from Heart Disease stored in UC Irvine Machine Learning Repository, a set of public datasets for scientific research [36]. The datasets contains 13 attribute columns and one column of the predicted attribute to judge the prevalence of Heart Disease. The third dataset comes from diabetes-related data set published by NHANES [37] in 2015-2016. The purpose of this dataset is to develop a method to calculate the risk of diabetes mellitus, which contains 11 attribute columns and a judgment column.

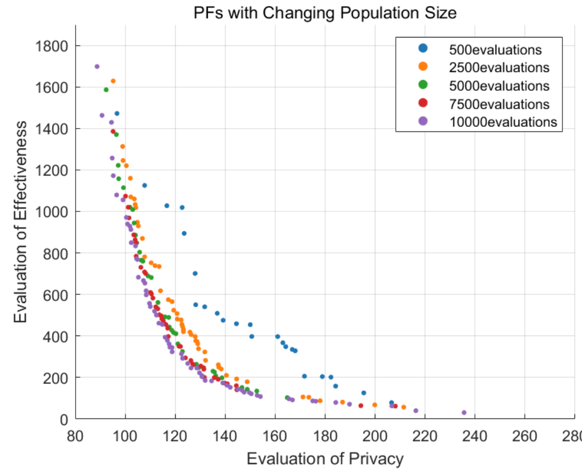


Fig. 4: PFs obtained by ESA

B. Presentation of FDPEL

In the first part, the dataset we use is the Mother's Significant Feature (MSF) dataset. After screening and data cleaning, we selected 89 attribute columns and one decision column, and digitized them for data processing.

Different from other MOPs, differential privacy is a noise-adding algorithm based on probability distribution, which is random. Even if the same privacy budget is set, the evaluation of the privacy and validity of the data set will be slightly biased. There is a special case: the same set of parameters will also produce a dominant relationship. With the increase of population generations, this situation will occur with great probability. Therefore, for privacy budget sets E_1 and E_2, if there is a dominant relationship, we can not simply describe it as that E_1's performance is completely superior to E_2's, but as " E_1's potential to generate a better solution set is superior to E_2's ".

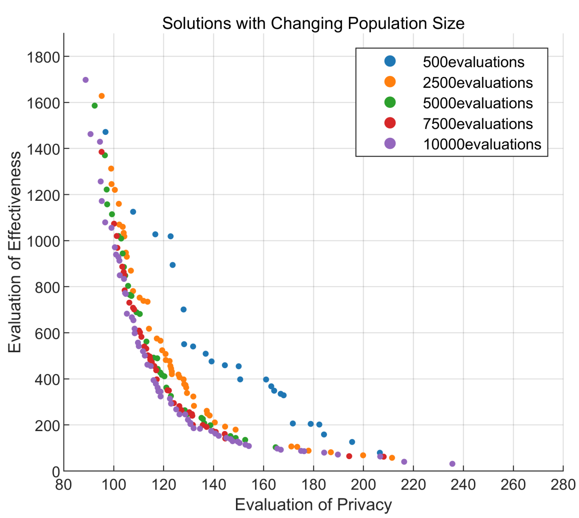


Fig. 3: All Solutions Obtained by ESA

Fig.3 and Fig.4 respectively show all solutions and PFs obtained by running ESA. We set a generation to produce 100 populations and show the results when the total population changes from 500 to 50000. The x axis represents the privacy performance of the dataset after differential privacy, and the smaller the value, the better the privacy protection effect. The y axis represents the validity performance of the dataset after differential privacy, and the smaller the data, the smaller the influence of differential privacy on the effectiveness. We can see the superior performance of ESA. At the initial stage of the algorithm's operation, it achieved accelerated convergence based on ALMOEA framework, and achieved superior performance in 2500 generations. By comparing PFs when total population changes from 500 to 50000, we can find that the diversity of the population has also been expanded and maintained.

TABLE II: Partial data before differential privacy

ε	**1.53**	**1.28**	**1.85**	**1.09**
ID	Age of Mother	Weight before Preg	Wt before Delivery	Yrs of Marriage
1	29	**59→58**	156	0
2	24	54	145	0
3	28	**62→61**	151	0
4	25	49	151	0
5	**21→22**	39	151	0
6	**32→31**	56	156	0
7	23	40	141	0
8	23	52	**159→160**	0
9	29	**59→58**	149	0
10	28	69	156	0
11	25	51	145	0
12	30	75	156	1
13	22	40	142	**0→1**
14	26	57	154	0
15	25	50	144	0
16	27	60	151	0
17	29	62	150	1
18	24	45	131	0
19	37	82	171	0
20	33	69	169	0

Table.II shows the partial data before and after applying differential privacy, along with the corresponding privacy budgets. Because of the large amount of data, we selected some representative columns, which contain some continuous values, such as Age of Mother etc., and a discrete values,

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3366889

Yrs of Marriage. It can be seen that due to the different privacy budgets, the data is disturbed differently. Our privacy and effectiveness evaluation is also based on changes of datasets. According to the Sec.III-A, we standardized the original information before differential privacy, ensuring that all indicators carry equal weight in the evaluation. For example, for the Weight before Pregnant attribute and Years of Marriage attribute in the Table.II, even if the privacy budget is the same, it seems that the data in the previous column has caused greater disturbance. However, the latter column only has two values of 0 and 1, so the information change degree from 0 to 1 is higher than that from 59 to 58 in the previous column. Therefore, the privacy budget directly reflects the change degree of information, rather than the absolute change value of disturbance.

In this experiment, we choose the population number of each generation as 100. The number of population produced in each generation can also be set to 50, 200 and other values. When the population number of each generation is set small, the iterative base of the population becomes smaller, and the diversity of the generated solution set decreases, resulting in poor performance. Therefore, in general, the number of populations will be set at a larger value. However, this will also lead to a problem: it is difficult to choose which solutions are more suitable through people's subjective judgment. We designed the method of double clustering to facilitate people to select the solution set.

The process and effect of double clustering are also important for results. Fig.5 shows the effect of the first reunion class. Based on the evaluation of privacy and validity, all solutions are divided into five clustering clusters, which are privacy-first, efficiency-first, privacy-focused, efficiency-focused and balance. The original huge solution set is divided into smaller populations for users to filter.

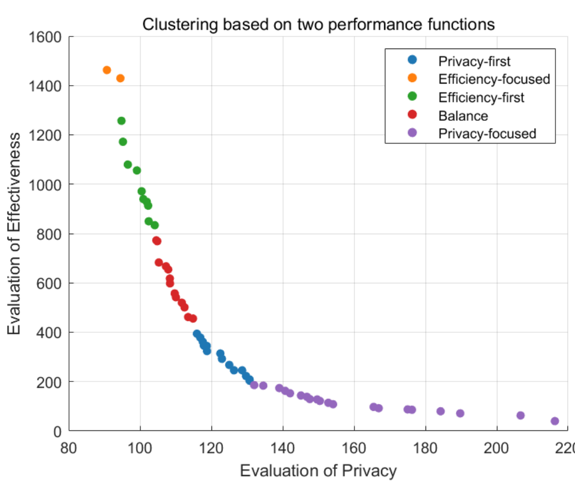


Fig. 5: Clustering Based on Two Performance Functions

Fig.6 shows the effect of the second clustering. The second clustering is based on the solution (privacy budget), so that users can filter according to their preferences. The x axis represents 89 attributes, and the y axis represents the value of privacy budget. Each line represents the average value of the divided clusters, and the characteristics of each cluster can be clearly observed through the image.

Table.III shows the results of double clustering. For example, we prefer to choose the population with Privacy-focused characteristics, and we only need to screen it separately, and at the same time, the results of the second clustering will be displayed on the right, as the bold part in the Table.III. Users can combine Fig.5 and Fig.6 to make more accurate choices and conduct differential privacy treatment on their own data.

C. Performance Analysis of FDPEL

In this section, we analyze the computational complexity and cost, and conduct ablation experiments to verify the effectiveness of the evaluation index.

We analyze the computational complexity and overhead of FDPEL. Obviously, it can be analyzed from four aspects: differential privacy, evaluation system, ESA and double clustering. In the process of differential privacy, the computational complexity $O(mN)$ of adding noise separately for each column of size m and $O(mNn)$ for n columns, N is the population size in one round. In evaluation system, the computational complexity of f_{11}, f_{12}, f_{13} of privacy evaluation is $O(Nn)$, $O(mNn)$ and $O(mNn)$ respectively, and the computational complexity of f_{21}, f_{22}, f_{23} of validity evaluation is $O(mNn + fmN)$, $O(mNn)$ and $O(mNn)$ respectively, where f is the number of discrete values in the judgment column. Since $f \leq n$, $O(fmN)$ can be ignored. In ESA, it should be analyzed separately from two stages. The computational complexity of the accelerated convergence stage includes three aspects: the process of training to generate mlp is $O(aN^2 + Nnk)$, the process of reproduction is $O(Nnk)$, and the process of environment selection is $O(aN^2)$, where a is the target number, N is the population size, n is the number of variables, and k is the number of hidden neurons, which is generally set to 10. The stage of expanding diversity is relatively simple, and the time complexity is $O(aN^2)$. In the dual clustering stage, the time complexity of the first reunion class is $O(tpaN)$, t is the number of iterations, p is the number of clusters, N is the population size, and a is the target number. The time complexity of the second clustering depends on the distribution of points, and the worst case is $O(N^2)$, where N is the population size. Based on the above analysis, the worst computational complexity and runtime of FDPEL are shown in Table.IV. The specific parameters are set as m = 450, n = 89, N = 100, f = 2, a = 2, k = 10, t = 100 and p = 5. The runtime of the first three parts in the table is one round, while the runtime of double clustering is once after the optimization is completed. All the solvers were run on a personal computer having a AMD Ryzen5 5600H CPU, 3.30GHz (processor), and 16 GB(RAM).

TABLE IV: The worst computational complexity of FDPEL

DP	Evaluation Metrics	ESA	Double Clustering	
			first	second
$O(mNn)$	$O(mNn)$	$O(mN^2 + Nnk)$	$O(tkmn)$	$O(N^2)$
$0.38s$	$6.81s$	$6.91s$	$0.07s$	$0.13s$

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3366889

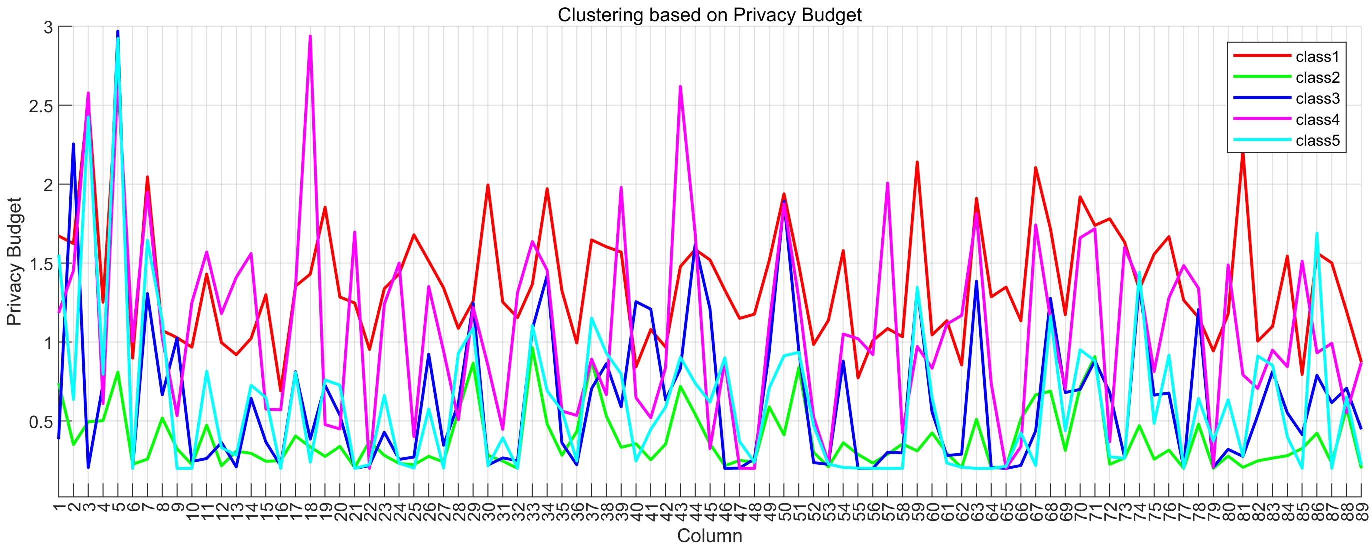


Fig. 6: Clustering Based on Privacy Budget

TABLE III: All Results of Double Clustering

ID	First clustering	Second clustering	ID	First clustering	Second clustering	ID	First clustering	Second clustering
1	Efficiency-first	single point	21	Efficiency-first	class3	41	**Privacy-first**	**single point**
2	Privacy-focused	class1	22	Balance	class1	42	Efficiency-focused	class1
3	Efficiency-first	single point	23	Balance	class1	43	Efficiency-first	single point
4	Efficiency-first	single point	24	Efficiency-first	single point	44	Efficiency-first	single point
5	Privacy-focused	class1	25	Privacy-focused	class1	45	**Privacy-first**	**class5**
6	Efficiency-first	single point	26	Efficiency-first	class3	46	Balance	single point
7	Efficiency-first	single point	27	**Privacy-first**	**class4**	47	Efficiency-first	class5
8	Efficiency-first	single point	28	Efficiency-focused	class1	48	**Privacy-first**	**single point**
9	Privacy-focused	class1	29	Balance	single point	49	Balance	single point
10	Efficiency-first	single point	30	Efficiency-focused	class1	50	Efficiency-first	single point
11	Efficiency-first	single point	31	Efficiency-focused	class1	51	**Privacy-first**	**single point**
12	Efficiency-focused	class1	32	Balance	class1	52	**Privacy-first**	**class2**
13	Efficiency-first	single point	33	**Privacy-first**	**class4**	53	**Privacy-first**	**class5**
14	Balance	class1	34	Balance	class1	54	Efficiency-first	single point
15	Efficiency-first	single point	35	Efficiency-focused	class1	55	Efficiency-focused	class1
16	**Privacy-first**	**class2**	36	Balance	class1	56	Efficiency-focused	class1
17	Efficiency-focused	class1	37	Efficiency-first	single point	57	**Privacy-first**	**class2**
18	Efficiency-focused	class1	38	Efficiency-first	single point	58	Efficiency-first	single point
19	**Privacy-first**	**single point**	39	**Privacy-first**	**class4**	59	**Privacy-first**	**single point**
20	Efficiency-focused	class1	40	**Privacy-first**	**single point**	60	Balance	class4

We conducted ablation experiments on FDPEL to verify the effectiveness of the evaluation index. Under the condition of the same parameters, each index is removed in turn, and the optimization results are evaluated to evaluate the performance of the generated PFs. We use two indicators IGD and HV, and the experimental results are shown in Table.V and Fig.ablation, where base refers to the case without ablation, and from f_{11} to f_{13} represent the first item of privacy function to the third item of validity function, as used in Subsection.III-B. It can be clearly seen that removing each index has great influence on the evaluation system.

TABLE V: Ablation experiments of FDPEL

	Base	-f_{11}	-f_{12}	-f_{13}	$-f_{21}$	-f_{22}	-f_{23}
IGD($e2$)	**1.04**	2.35	2.17	2.24	2.24	2.23	2.24
HV($e5$)	**4.61**	1.34	1.80	1.81	1.78	1.79	1.82

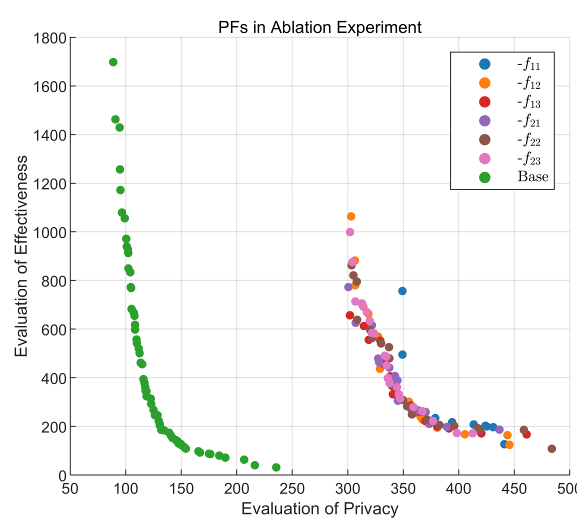


Fig. 7: PFs in Ablation Experiment

D. Contrast Experiment of ESA

In this section, we compare ESA with other multi-objective optimization algorithms to show the superiority of our designed algorithm. When the algorithms is running, the parameters are set as follows: population size is 100 and maxFE is 10000, which are the most widely used in multi-objective algorithm effect verification. Fig.8 shows the final generated PF, in which the non-dominant solution is excluded. It can be clearly seen that the performance and diversity of ESA is better than other MOEA.

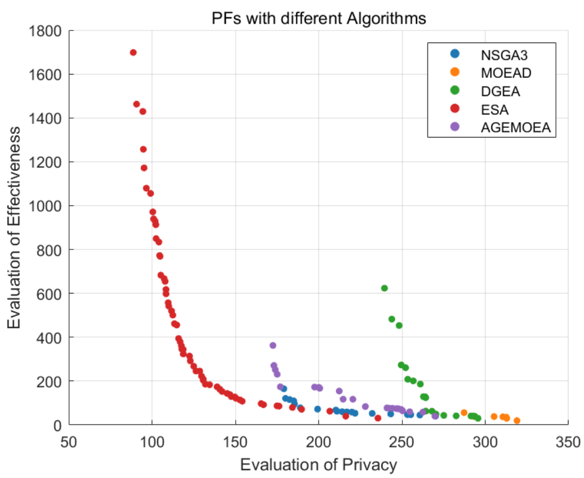


Fig. 8: PFs with Different Algorithms

We compared the evaluation methods commonly used in MOEA, such as IGD, HV and Set Coverage, and found that ESA performed very well, as shown in Fig.10.

Fig.10a shows the IGD performance of different MOEA algorithms. It can also be clearly observed that the IGD of ESA drops rapidly and finally reaches a better value balance. It should be noted here that because the pareto optimal frontier of this problem is unknown, we construct a simulated pareto optimal frontier [38]. The concrete construction method is to draw a right-angled polyline according to the optimal solution of single-objective optimization of four algorithms. The x axis of the point with the smallest privacy evaluation in Pareto curve represents the optimal solution of single-objective optimization for the first objective function. Similarly, the y axis of the point with the smallest effectiveness evaluation indicates the optimal solution for the single-objective optimization of the second objective function, as shown in Fig.9.

Fig.10b shows the HV performance of different algorithms. The comparison of HV is a commonly used evaluation method without finding pareto optimal frontier. It can be observed that the HV value of ESA rises rapidly and tends to be stable gradually. It can be verified that the diversity and comprehensive performance of PF generated by ESA are higher than other algorithms.

Fig.10c shows the Set Coverage(SC) performance of PF produced by ESA compared with PF generated by other algorithms. Set Coverage is used to evaluate the dominance of two PFs. Assuming that A and B are two PF, then SC(A,B) can be expressed as

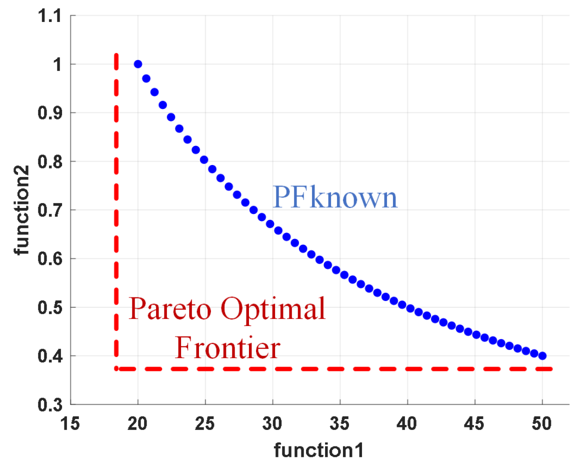


Fig. 9: Simulated Pareto Optimal Frontier

$$SC(A,B) = \frac{|\{b \in B | \exists a \in A : a \succ b\}|}{|B|} \tag{25}$$

$|B|$ indicates the number of solutions in B, and $C(A,B)$ indicates the percentage that the solution in B is dominated by a solution of A. The greater the value of $C(A,B)$, the better the performance of A,B. It can be easily seen that the PF solution set generated by ESA dominates PF generated by other algorithms to some extent, and the value of SC is close to the maximum value of 1.

E. General Applicability of FDPEL

FDPEL has strong versatility and adaptability in health-related datasets, and users can easily apply FDPEL to the datasets they want to publish. It is not only suitable for privacy protection of larger-scale medical datasets, but also suitable for smaller datasets.

In this subsection, we try to apply FDPEL to datasets of other health topics to test its universality. It mainly shows its performance on two datasets.

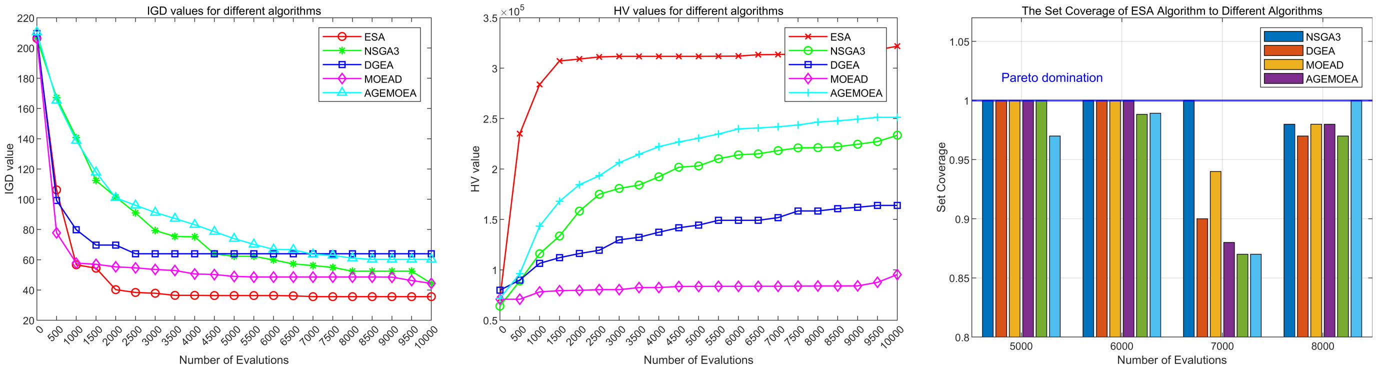
The first data set comes from Heart Disease stored in UC Irvine Machine Learning Repository, which is mentioned in SubsectionIV-A. The generated PF and the result of double clustering are shown in Fig.11 and Fig.12.

The second dataset is diabetes-related, which is mentioned in SubsectionIV-A. The pareto curve generated by it and the result of double clustering are shown in Fig.13 and Fig.14.

F. Discussion

In the experiment, we verified the superior performance and versatility of FDPEL. First of all, we showed FDPEL and analysed its performance. Secondly, we compare ESA algorithm with other evolutionary algorithms, and show its optimization performance through some performance indicators. Thirdly, we extend FDPEL to the other two dataset, showing its universality. Of course, there are still some shortcomings in the design of the experiment. We have not applied FDPEL to large-scale data sets (the number of attributes is greater than 1000), which will be our next research direction.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3366889



(a) IGD Values for Different Algorithms (b) HV Values for Different Algorithms (c) The Value of Set Coverage(ESA, other algorithm)

Fig. 10: Contrast Experiment of ESA

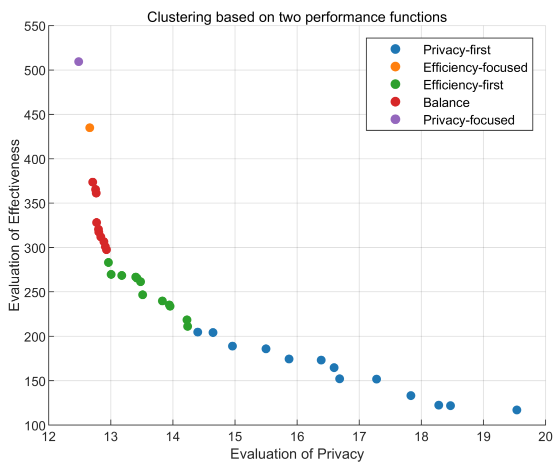


Fig. 11: FDPEL on the Heater Disease Dataset(1)

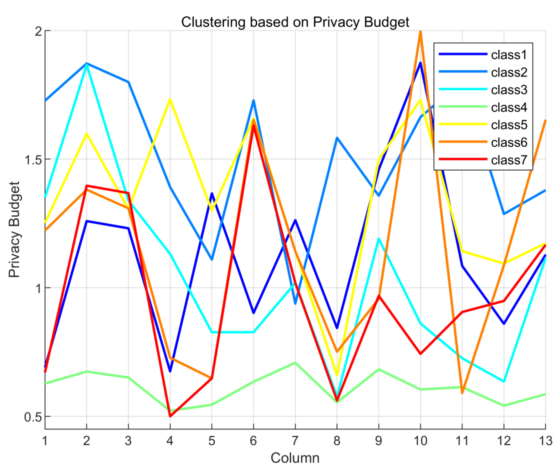


Fig. 12: FDPEL on the Heater Disease Dataset(2)

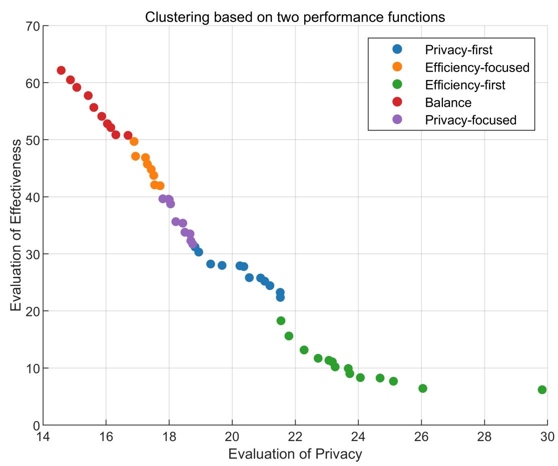


Fig. 13: FDPEL on the Diabetes Dataset(1)

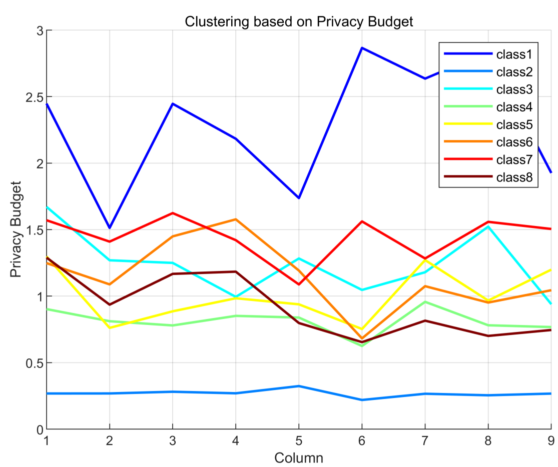


Fig. 14: FDPEL on the Diabetes Dataset(2)

V. Conclusion

With the rapid development of Internet of Medical Things (IOMT), lots of valuable and private medical data need to be protected. We propose Flexible Differential Privacy Algorithm based on Evolutionary Learning (FDPEL), which realizes the privacy protection of medical data of different scales. FDPEL consists of three parts: Firstly, noise disturbance of medical data using differential privacy. Secondly, Environment Switching Algorithm (ESA) based on evolutionary learning is used to adjust privacy budgets of different attributes and balance data privacy and data validity. ESA has excellent performance, which can ensure convergence speed and optimization performance at the same time. Thirdly, A double clustering method is used to select the appropriate solution from the huge PF. Finally, we verify the superior performance and versatility of FDPEL through simulation experiments. FDPEL can be easily migrated to IMOT of various scales for privacy protection.

REFERENCES

[1] D. Tian, N. Yu, J. Yu, H. Zhang, J. Sun, and X. Bai, "Research on dual-line array subpixel scanning imaging for iomt-based blood cell analysis system," *IEEE Internet of Things Journal*, vol. 10, no. 1, pp. 367–377, 2023.

[2] T. Wei, S. Liu, and X. Du, "Learning-based efficient sparse sensing and recovery for privacy-aware iomt," *IEEE Internet of Things Journal*, vol. 9, no. 12, pp. 9948–9959, 2022.

[3] X. Liu, J. Zhao, J. Li, B. Cao, and Z. Lv, "Federated neural architecture search for medical data security," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5628–5636, 2022.

[4] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2021.

[5] M. Kumar, Kavita, S. Verma, A. Kumar, M. F. Ijaz, and D. B. Rawat, "Anaf-iomt: A novel architectural framework for iomt-enabled smart healthcare system by enhancing security based on recc-vc," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 12, pp. 8936–8943, 2022.

[6] A. Ghubaish, T. Salman, M. Zolanvari, D. Unal, A. Al-Ali, and R. Jain, "Recent advances in the internet-of-medical-things (iomt) systems security," *IEEE Internet of Things Journal*, vol. 8, no. 11, pp. 8707–8718, 2021.

[7] R. Gupta, I. Gupta, A. K. Singh, D. Saxena, and C.-N. Lee, "An iot-centric data protection method for preserving security and privacy in cloud," *IEEE Systems Journal*, vol. 17, no. 2, pp. 2445–2454, 2023.

[8] X. Wang, J. Hu, H. Lin, W. Liu, H. Moon, and M. J. Piran, "Federated learning-empowered disease diagnosis mechanism in the internet of medical things: From the privacy-preservation perspective," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 7, pp. 7905–7913, 2023.

[9] T. Murakami and Y. Sei, "Automatic tuning of privacy budgets in input-discriminative local differential privacy," *IEEE Internet of Things Journal*, vol. 10, no. 18, pp. 15 990–16 005, 2023.

[10] H. Huang, D. Zhang, F. Xiao, K. Wang, J. Gu, and R. Wang, "Privacy-preserving approach pbcn in social network with differential privacy," *IEEE Transactions on Network and Service Management*, vol. 17, no. 2, pp. 931–945, 2020.

[11] M. Zhang, Y. Chen, and J. Lin, "A privacy-preserving optimization of neighborhood-based recommendation for medical-aided diagnosis and treatment," *IEEE Internet of Things Journal*, vol. 8, no. 13, pp. 10 830–10 842, 2021.

[12] K. Qiao, J. Liang, K. Yu, M. Wang, B. Qu, C. Yue, and Y. Guo, "A self-adaptive evolutionary multi-task based constrained multi-objective evolutionary algorithm," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 7, no. 4, pp. 1098–1112, 2023.

[13] L. Zhang, H. Zhang, S. Liu, C. Wang, and H. Zhao, "A community division-based evolutionary algorithm for large-scale multi-objective recommendations," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 7, no. 5, pp. 1470–1483, 2023.

[14] L. Wu, C. Qin, Z. Xu, Y. Guan, and R. Lu, "Tcpp: Achieving privacy-preserving trajectory correlation with differential privacy," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 4006–4020, 2023.

[15] X. Ye, Y. Zhu, M. Zhang, and H. Deng, "Differential privacy data release scheme using microaggregation with conditional feature selection," *IEEE Internet of Things Journal*, vol. 10, no. 20, pp. 18 302–18 314, 2023.

[16] B. Jiang, J. Li, H. Wang, and H. Song, "Privacy-preserving federated learning for industrial edge computing via hybrid differential privacy and adaptive compression," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 1136–1144, 2023.

[17] H. Jiang, J. Pei, D. Yu, J. Yu, B. Gong, and X. Cheng, "Applications of differential privacy in social network analysis: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 108–127, 2023.

[18] M. Yang, I. Tjuawinata, K. Y. Lam, J. Zhao, and L. Sun, "Secure hot path crowdsourcing with local differential privacy under fog computing architecture," *IEEE Transactions on Services Computing*, vol. 15, no. 4, pp. 2188–2201, 2022.

[19] V. Ravi, T. D. Pham, and M. Alazab, "Attention-based multidimensional deep learning approach for cross-architecture iomt malware detection and classification in healthcare cyber-physical systems," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 4, pp. 1597–1606, 2023.

[20] Z. Lian, Q. Zeng, W. Wang, T. R. Gadekallu, and C. Su, "Blockchain-based two-stage federated learning with non-iid data in iomt system," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 4, pp. 1701–1710, 2023.

[21] X. Jia, M. Luo, H. Wang, J. Shen, and D. He, "A blockchain-assisted privacy-aware authentication scheme for internet of medical things," *IEEE Internet of Things Journal*, vol. 9, no. 21, pp. 21 838–21 850, 2022.

[22] P. Zeng, Z. Zhang, R. Lu, and K.-K. R. Choo, "Efficient policy-hiding and large universe attribute-based encryption with public traceability for internet of medical things," *IEEE Internet of Things Journal*, vol. 8, no. 13, pp. 10 963–10 972, 2021.

[23] R. Gupta, D. Saxena, I. Gupta, and A. K. Singh, "Differential and triphase adaptive learning-based privacy-preserving model for medical data in cloud environment," *IEEE Networking Letters*, vol. 4, no. 4, pp. 217–221, 2022.

[24] J. H. Zheng, Y. N. Kou, Z. X. Jing, and Q. H. Wu, "Towards many-objective optimization: Objective analysis, multi-objective optimization and decision-making," *IEEE Access*, vol. 7, pp. 93 742–93 751, 2019.

[25] Z. Zhou, S. Li, G. Qin, M. Folkert, S. Jiang, and J. Wang, "Multi-objective-based radiomic feature selection for lesion malignancy classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 1, pp. 194–204, 2020.

[26] Y. Sun, G. G. Yen, and Z. Yi, "Igd indicator-based evolutionary algorithm for many-objective optimization problems," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 2, pp. 173–187, 2019.

[27] J. Ding, S. Dauzre-Prs, L. Shen, and Z. L, "A novel evolutionary algorithm for energy-efficient scheduling in flexible job shops," *IEEE Transactions on Evolutionary Computation*, vol. 27, no. 5, pp. 1470–1484, 2023.

[28] S. Liu, Q. Lin, L. Feng, K.-C. Wong, and K. C. Tan, "Evolutionary multitasking for large-scale multiobjective optimization," *IEEE Transactions on Evolutionary Computation*, vol. 27, no. 4, pp. 863–877, 2023.

[29] S. Liu, Q. Lin, Q. Li, and K. C. Tan, "A comprehensive competitive swarm optimizer for large-scale multiobjective optimization," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 9, pp. 5829–5842, 2022.

[30] Y. Xu, C. Xu, H. Zhang, L. Huang, Y. Liu, Y. Nojima, and X. Zeng, "A multi-population multi-objective evolutionary algorithm based on the contribution of decision variables to objectives for large-scale multi/many-objective optimization," *IEEE Transactions on Cybernetics*, vol. 53, no. 11, pp. 6998–7007, 2023.

[31] S. Liu, J. Li, Q. Lin, Y. Tian, and K. C. Tan, "Learning to accelerate evolutionary search for large-scale multiobjective optimization," *IEEE Transactions on Evolutionary Computation*, vol. 27, no. 1, pp. 67–81, 2023.

[32] Y. Hua, H. Zhu, and Y. Xu, "Multi-objective optimization design of bearingless permanent magnet synchronous generator," *IEEE Transactions on Applied Superconductivity*, vol. 30, no. 4, pp. 1–5, 2020.

[33] X. Xie, J. Zheng, M. Feng, S. He, and Z. Lin, "Multi-objective mayfly optimization algorithm based on dimensional swap variation for rfid network planning," *IEEE Sensors Journal*, vol. 22, no. 7, pp. 7311–7323, 2022.

[34] Y. Xu, J. Chen, S. Huang, K. Xu, C. Lu, and P. Li, "Multi-objective optimization design of repetitive pulse magnetic field system," *IEEE Transactions on Applied Superconductivity*, vol. 30, no. 4, pp. 1–6, 2020.

[35] G. Marvin, "Maternal health medical advice," 2023. [Online]. Available: https://dx.doi.org/10.21227/fb4h-zr30

[36] S. W. P. M. Janosi, Andras and R. Detrano, "Heart Disease," UCI Machine Learning Repository, 1988, DOI: https://doi.org/10.24432/C52P4X.

[37] Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS), "National health and nutrition examination survey data," Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2023. [Online]. Available: https://www.cdc.gov/nchs/nhanes/index.htm

[38] Q. Kang, S. Feng, M. Zhou, A. C. Ammari, and K. Sedraoui, "Optimal load scheduling of plug-in hybrid electric vehicles via weight-

aggregation multi-objective evolutionary algorithms," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 9, pp. 2557–2568, 2017.

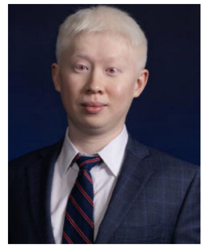
Yongxiang Kuang is currently a student with the College of Oceanography and Space Informatics, China University of Petroleum (East China), majoring in communication engineering. He won the National Scholarship of the Ministry of Education of China in 2022 and 2023. His major research interests include internet of underwater things, underwater information processing, privacy-preserving, and multi-objective optimization.

Bin Jiang (Member, IEEE) received the Ph.D. degree in information and communication engineering from Tianjin University, Tianjin, China, in 2020, where he also received the B.S. and M.S. degree in 2013 and 2016. He is currently an Associate Professor with the College of Oceanography and Space Informatics, China University of Petroleum (East China). Dr.Jiang was a joint Ph.D student in the Security and Optimization for Networked Globe Laboratory, Department of Electrical Engineering and Computer Science, Embry-Riddle Aeronautical University, Daytona Beach, FL, USA, from Nov. 2017 to Feb.2018, and from Oct.2018 to Oct.2019. He was a post-doctoral in College of Computer Science and Software Engineering, Shenzhen University, China, from Aug.2020 to July.2022. His major research interests include Internet of Things, underwater information processing, communications and security.

Xuerong Cui (Member, IEEE) received the Ph.D. degree in Information Science and Engineering from the Ocean University of China in 2012 and the Master degree in Computer Application Technology from China University of Petroleum in 2003. He joined the China University of Petroleum in 2003 in the Department of Computer and Communication Engineering, and from 2019, he works in the College of Oceanography and Space Informatics. From Feb. 2015 to Feb. 2016 he worked in the University of Victoria as a Visiting Scholar. His research interests include underwater communication, detection, and navigation, big data and artificial intelligence, wireless positioning and navigation, *etc.*

ShiBao Li (Member, IEEE) was born in 1978. He received the B.S. and M.S. degrees in computer science from the University of Petroleum, Qingdao, China, in 2002 and 2009, respectively.

He is currently a Professor with the China University of Petroleum, Qingdao. His research interests include indoor localization technology, the Internet of Things, wireless networks, and mobile computing.



Yongxin Liu (Senior Member, IEEE) is an assistant professor of data science at Embry-Riddle Aeronautical University (ERAU). His research focuses on Explainable AI for aviation, Unmanned Aerial Systems and Internet of Things. He's the creator of zero-bias deep learning, which is applied in explainable AI and anomaly detection. He's also the creator of the Channel Separation Incremental Learning algorithm, which harnesses the catastrophic forgetting issues in deep learning with mathematical assurance and applied in ADS-B signal identification systems. Other than that, his early work in unauthorized drone detection has resulted in two patents and was cited as first reference by the Department of Defense in August 2021. Dr. Liu has received 4 best papers awards and the 2021 Harry Rowe Mimno award from IEEE Aerospace and Electronics Systems Society. His recent research project is funded under USDOT Tier 1 UTC Transportation Cybersecurity Center for Advanced Research and Education (CYBER-CARE).

Houbing Song (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, VA, in August 2012.

He is currently a Tenured Associate Professor, the Director of the NSF Center for Aviation Big Data Analytics (Planning), the Associate Director for Leadership of the DOT Transportation Cybersecurity Center for Advanced Research and Education (Tier 1 Center), and the Director of the Security and Optimization for Networked Globe Laboratory (SONG Lab, www.SONGLab.us), University of Maryland, Baltimore County (UMBC), Baltimore, MD. Prior to joining UMBC, he was a Tenured Associate Professor of Electrical Engineering and Computer Science at Embry-Riddle Aeronautical University, Daytona Beach, FL. He serves as an Associate Editor for IEEE Transactions on Artificial Intelligence (TAI) (2023-present), IEEE Internet of Things Journal (2020-present), IEEE Transactions on Intelligent Transportation Systems (2021-present), and IEEE Journal on Miniaturization for Air and Space Systems (J-MASS) (2020-present). He was an Associate Technical Editor for IEEE Communications Magazine (2017-2020). He is the editor of eight books, the author of more than 100 articles and the inventor of 2 patents. His research interests include cyber-physical systems/internet of things, cybersecurity and privacy, and AI/machine learning/big data analytics. His research has been sponsored by federal agencies (including National Science Foundation, National Aeronautics and Space Administration, US Department of Transportation, and Federal Aviation Administration, among others) and industry. His research has been featured by popular news media outlets, including IEEE GlobalSpec's Engineering360, Association for Uncrewed Vehicle Systems International (AUVSI), Security Magazine, CXOTech Magazine, Fox News, U.S. News & World Report, The Washington Times, and New Atlas.

Dr. Song is an IEEE Fellow (for contributions to big data analytics and integration of AI with Internet of Things), and an ACM Distinguished Member (for outstanding scientific contributions to computing). He is an ACM Distinguished Speaker (2020-present), an IEEE Vehicular Technology Society (VTS) Distinguished Lecturer (2023-present) and an IEEE Systems Council Distinguished Lecturer (2023-present). Dr. Song has been a Highly Cited Researcher identified by Clarivate (2021, 2022) and a Top 1000 Computer Scientist identified by Research.com. He received Research.com Rising Star of Science Award in 2022 (World Ranking: 82; US Ranking: 16). In addition to 2021 Harry Rowe Mimno Award bestowed by IEEE Aerospace and Electronic Systems Society, Dr. Song was a recipient of 10+ Best Paper Awards from major international conferences, including IEEE CPSCom-2019, IEEE ICII 2019, IEEE/AIAA ICNS 2019, IEEE CBDCom 2020, WASA 2020, AIAA/ IEEE DASC 2021, IEEE GLOBECOM 2021 and IEEE INFOCOM 2022.