4-2019

CINEMA: Efficient and privacy-preserving online medical primary diagnosis with skyline query

Jianfeng HUA

Hui ZHU

Fengwei WANG

Ximeng LIU
Singapore Management University, xmliu@smu.edu.sg

Rongxing LU

See next page for additional authors

Citation

HUA, Jianfeng; ZHU, Hui; WANG, Fengwei; LIU, Ximeng; LU, Rongxing; LI, Hao; and ZHANG, Yeping. CINEMA: Efficient and privacy-preserving online medical primary diagnosis with skyline query. (2019). *IEEE Internet of Things*. 6, (2), 1450-1461.
Available at: https://ink.library.smu.edu.sg/sis_research/5151

Published in IEEE Internet of Things Journal, 2019 April, 6 (2), 1450-1461.
https://doi.org/10.1109/JIOT.2018.2834156

CINEMA: Efficient and Privacy-Preserving Online Medical Primary Diagnosis With Skyline Query

Jiafeng Hua, Hui Zhu, *Member, IEEE*, Fengwei Wang, Ximeng Liu, *Member, IEEE*, Rongxing Lu, *Senior Member, IEEE*, Hao Li, and Yeping Zhang

***Abstract*—Online medical primary diagnosis system, which can provide convenient medical decision support through applying mobile communication and data analysis technology, has been considered as a promising approach to improve the quality of healthcare service. However, it still faces many severe challenges on the privacy of users' health information and the accuracy of diagnosis result, which deter the wide adoption of online medical primary diagnosis system. In this paper, we propose an efficient and privacy-preserving online medical primary diagnosis (CINEMA) framework. Within CINEMA framework, users can access online medical primary diagnosing service accurately without divulging their medical data. Specifically, based on fast secure permutation and comparison technique, the encrypted user's query is directly operated at the service provider (SP) without decryption, and the diagnosis result can only be decrypted by the user, meanwhile, the diagnosis model in SP can also be protected. Through extensive analysis, we show that CINEMA can ensure that user's health information and healthcare SP's diagnosis model are kept confidential, and has significantly reduce computation and communication overhead. In addition, performance evaluations via implementing CINEMA demonstrate its effectiveness in term of the real environment.**

***Index Terms*—Efficiency, medical primary diagnosis, privacy-preserving, skyline computation.**

I. Introduction

With the proliferation of mobile communication technologies, collecting the various physiological data of a user by wearable devices anywhere and anytime has been in practice. Online medical primary diagnosis, which forms a critical capability to link physiological data with medical knowledge, has already shown great potential in improving the quality of healthcare services while lowering the healthcare cost [1]–[3]. As shown in Fig. 1, in online medical primary diagnosis system, a user can deploy portable sensors around body to collect various physiological data, such as electrocardiogram, blood pressure (BP), peripheral oxygen saturation, and blood glucose, even bioactive elements like DNA, viruses, enzymes, and cell receptors [4]. These physiological data is sent to the healthcare service provider (SP) by smart terminals, which run various medical applications to receive real-time medical instruction from the healthcare SP.

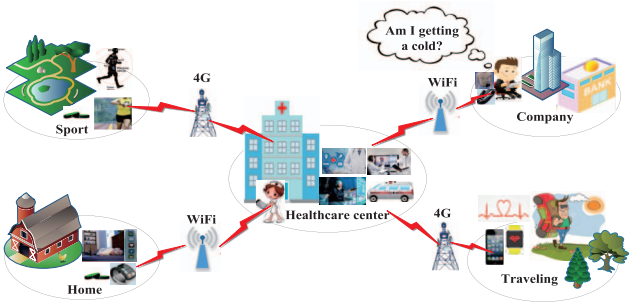


Fig. 1. Online medical primary diagnosis system.

As a powerful tool for multicriteria data analysis, data mining, and decision making [5], skyline query returns a set of interesting points which are the best tradeoffs between the different dimensions of a huge data space [6], [7]. By querying the points which are as good or better in all dimensions and better in at least one dimension, skyline query has been received significant attention on centralized database [8], [9], distributed database [10]–[12], and similarity search [13], [14]. However, the formal work mainly aims at finding the data points which are not worse than the other data points under the given query requirements, while not aware how to apply skyline query in medical diagnosis system to perform range query. Furthermore, the physiological data has a standard reference region and extremely rich in information with high dimension [15], and the single aggregated distance metric with all dimensions is always hard to define, which may be quite appropriate for skyline query applicated in medical data decision making [14].

Manuscript received January 28, 2018; revised April 16, 2018; accepted May 3, 2018. Date of publication May 8, 2018; date of current version May 8, 2019. The work of H. Zhu was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0802201, in part by the National Natural Science Foundation of China under Grant 61672411, Grant U1401251, and Grant 81600574, in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2016ZDJC-04 and Grant 2017JQ8001, and in part by the China 111 Project under Grant B16037. The work of R. Lu was supported in part by Natural Sciences and Engineering Research Discovery under Grant Rgpin 04009, in part by NBIF Start-Up under Grant Rif 2017-915012, in part by URF under Grant Nf-2017-05, and in part by HMF under Grant 2017 Ys-4. *(Corresponding author: Hui Zhu.)*

J. Hua, H. Zhu, F. Wang, and Y. Zhang are with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: jfhua@stu.xidian.edu.cn; zhuhui@xidian.edu.cn; xdwangfengwei@gmail.com; canhongping@gmail.com).

X. Liu is with the School of Information Systems, Singapore Management University, Singapore 188065, and also with the College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China (e-mail: snbnix@gmail.com).

R. Lu is with the Faculty of Computer Science, University of New Brunswick, Fredericton, NB E3B 5A3, Canada (e-mail: rlu1@unb.ca).

H. Li is with First Affiliated Hospital, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: hao.li215@mail.xjtu.edu.cn).

Digital Object Identifier 10.1109/JIOT.2018.2834156

However, the sensitivity of medical data is extremely critical in terms of user's privacy, accidental data leakage may lead huge psychological harm to the user and even threaten the human life [16], which may be the main stumble in blocking the popularization of online medical primary diagnosis. Generally, users are reluctant to send their health information directly to untrusted third-party organization to obtain medical instruction. Meanwhile, the diagnosis system is also private and valuable asset, the third-party organization is also unwillness to reveal any information about it as well [17]–[19]. Therefore, how to protect the privacy of users' medical data and the confidentiality of primary diagnosis system is crucially. Traditional anonymization techniques such as *k*-anonymity [20] and *l*-diversity [21] may be not quite suitable for protecting the user's privacy, due to the user's query data always contain sensitive data, such as age, blood types, or even fingerprints and DNA profiles, which may be able to reidentify an individual user easily [22], [23]. On the other hand, differential privacy has become the de facto standard for privacy-preserving data analytics [24], [25], but these randomization approaches are often unsuitable for medical primary diagnosis, as they distort the data making it unusable for critical inferences, which may lead to misdiagnosis. Different homomorphic encryption techniques are introduced in the medical diagnosis system [26]–[28], but the overhead of computation would be a stumbling block in making this technology popularization in medical primary diagnosis system.

In this paper, aiming at these above challenges, a precise diagnosis model is first proposed by using skyline computation over the medical dataset. Then, due to security and privacy concerns, we propose an efficient and privacy-preserving online medical primary diagnosis (CINEMA) framework. Within CINEMA, users can achieve privacy-preserving medical primary diagnosis service themselves according to the diagnosis model stored at the SP. The SP provides online disease prediction service without revealing the disease diagnosis model. Specifically, the main contributions of this paper are as fourfold.

1) The proposed skyline diagnosis model is more precise. Different from traditional medical diagnosis model, which rely the physiological indicator's min/max value statistical analyzed from a precise medical dataset to provide medical diagnosis service, skyline diagnosis model is constructed by a lot of contour data generated from the medical dataset.
2) The proposed CINEMA is secure and privacy-preserving for users' medical data and diagnosis model's parameters. Within CINEMA, the encrypted user's query is directly operated at the SP without decryption, and the diagnosis result can only be decrypted by the user, meanwhile, the diagnosis model in SP can also be kept secret from the user during the process.
3) The proposed CINEMA can provide the online medical primary diagnosis service with high accuracy. Although the user's query request and skyline diagnosis model are blurred during the process, the accuracy of diagnosis result is not affected, and the final experiment results over real medical dataset show that CINEMA can achieve high accuracy.
4) The proposed CINEMA is efficient in terms of computation and communication overhead. We develop a custom simulator and implement CINEMA in a real environment, performance evaluation demonstrates that our proposed CINEMA can provide efficient online medical primary diagnosis service with low computation and communication overhead.

The remainder of this paper is organized as follows. We formalize the system model, security requirements, and identify our design goal in Section II. In Section III, we briefly review the skyline computation, Bilinear pairing, and 0-encoding method. Then, we introduce a novel diagnosis model by using skyline computation and propose our CINEMA framework in Section IV, followed by the security analysis and performance evaluation in Sections V and VI, respectively. We also review some related works in Section VII. Finally, we conclude this paper in Section VIII.

II. System Model, Security Requirement, and Design Goal

In this section, we formalize the system model, security requirements, and identify our design goal.

A. System Model

In this paper, we consider a system model consisted of two parts: 1) SP and 2) user, as shown in Fig. 2. The SP offers CINEMA service based on users' query, which contains a lot of sensitive medical data. The user is equipped with computer or smartphone, which collects physiological data from wearable devices. These data is delivered to the SP as a query to achieve online medical primary diagnosis service.

SP is always considered as an authorized data analysis organization (such as centralized servers in hospital), it owns a diagnosis model built upon existing clinical datasets, which are initially provided by hospitals. SP should provides high-quality online medical primary diagnosis service for registered users without revealing any sensitive information of the diagnosis model. In addition, although SP is a server with high performance in computation and storage, since thousands of users may access query services at the same time, the efficiency of computation and communication should also be considered.

The registered user can enjoy the privacy-preserving online medical diagnosis service from SP by sending a query request, which contains a lot of sensitive medical data collected from wearable devices by smart phones or computers. Before sending to the SP, the medical data is transformed into ciphertext, and operated without decryption during the diagnosis process to avoid privacy leakage. Moreover, due to the limitation of the terminal in computation and storage, in order to lower the energy cost, the encryption technique is required to be efficient and lightweight enough to adapt the resource constraint terminals.

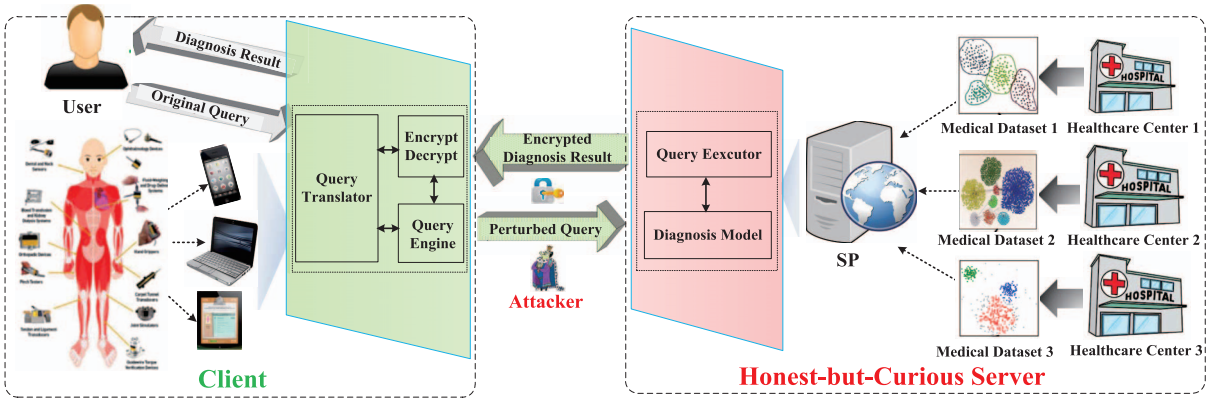


Fig. 2. System model under consideration.

B. Security Requirements

The privacy of users' medical query request and the confidentiality of diagnosis model are crucial for the success of online medical primary diagnosis service. In our system model, the users and SP are considered to be honest-but-curious. Specifically, the SP strictly executes the protocol specifications to guarantee the correctness of diagnosis service, but intends to extract the secret information about the user's medical data and diagnosis result; users honestly execute the operations to obtain the final diagnosis result, but they also try to gain knowledge about the diagnosis model; moreover, users may try to access the online medical primary diagnosis service without registering. Therefore, to guarantee the privacy of users' medical data and the confidentiality of diagnosis model, the following security requirements should be satisfied.

1) Privacy: Due to the query request contained lots of sensitive medical data, the user's query request should be protected from SP, i.e., even if SP receives the encrypted query request from the user, it cannot identify the user's sensitive medical data. Meanwhile, though SP computes the intermediate parameters according to the encrypted query request of user, it cannot obtain the final diagnosis result.

2) Confidentiality: The diagnosis model should be kept secret from the users since it is the intellectual property belongs to healthcare SP, i.e., even if the user obtains the intermediate parameters calculated by SP, she/he cannot identify the parameters of diagnosis model.

3) Authentication: Authentication an encrypted query request that is really sent by a legal user and has not been altered during the transmission, i.e., if an illegal user forges a query request, this malicious operation should be detected timely. Meanwhile, the responses from SP should also be authenticated so that the user can receive the authentic and reliable query result.

C. Design Goal

Based on the aforementioned system model and security requirements, our design goal is to develop an efficient and privacy-preserving online accurate medical primary diagnosis framework. Specifically, the following three objects should be achieved.

1) Security: The above-mentioned security requirements should be satisfied. According to the previous statement and analysis, without taking the security into consideration, the real application of the online medical primary diagnosis is far from in practice. Simultaneously, the confidentiality and authentication of the proposed framework should be achieved as well.

2) Accuracy: The accuracy of the diagnosis result should be guaranteed. In order to provide high-quality online medical primary diagnosis service, the designed privacy-preserving strategy cannot compromise the accuracy of diagnosis result. Therefore, the proposed framework should also achieve a high accuracy.

3) Efficiency: Low communication overhead and computation complexity should be guaranteed. Considering the real time requirements of online medical primary diagnosis service and the diversity of terminals, which might be constrained in computation and communication, the proposed framework should achieve high communication and computation efficiency.

III. Preliminaries

In this section, we first review the definition of skyline computation, which serves as the basis of our proposed framework, then we introduce the 0-encoding and 1-encoding method, bilinear pairing technique [29].

A. Skyline Definitions

We first introduce the definition of skyline computation [6].

Definition 1 (Skyline Computation): Given a dataset $P = \{P_1, \ldots, P_n\}$ in m-dimensional space. Considering that P_a and P_b are two different points in P, we say P_a dominated P_b, denoted by $\mathrm{dom}(P_a, P_b)$ if it satisfies the following conditions: 1) $\forall 1 \leq j \leq m$, $P_a[j] \leq P_b[j]$ and 2) at least there exists one j, $P_a[j] < P_b[j]$, where $P_i[j]$ is the jth dimension of P_i and $1 \leq i \leq n$. The skyline set is a set of points $\mathrm{Sky}(P) \subseteq P$ which are not dominated by any other points. The points in $\mathrm{Sky}(P)$ are called skyline point.

B. Bilinear Pairing

Let $\mathbb{G}$ and $\mathbb{G}_{\mathbb{T}}$ be two cyclic groups with the same prime order q, and g is a generator of group $\mathbb{G}$. Suppose $\mathbb{G}$ and $\mathbb{G}_{\mathbb{T}}$ are equipped with a pairing, i.e., a nondegenerated and efficiently computable bilinear map $\hat{e} : \mathbb{G} \times \mathbb{G} \longrightarrow \mathbb{G}_{\mathbb{T}}$ has the following properties.

1) Bilinearity: $\forall g, h \in \mathbb{G}$, and $\forall a, b \in \mathbb{Z}_q$, we have $\hat{e}(g^a, h^b) = \hat{e}(g, h)^{ab}$.

2) Nondegeneracy: $\exists$ at least one g, h, where $g, h \in \mathbb{G}$, which satisfies the condition that $\hat{e}(g, h) \neq 1_{\mathbb{G}_{\mathbb{T}}}$.

3) Computable: $\forall g, h \in \mathbb{G}$, there is an efficient algorithm to compute $\hat{e}(g, h)$.

Definition 2: A bilinear parameter generator *Gen* is a probabilistic algorithm that takes a security parameter k as input, and outputs a five-tuple $(q, g, \mathbb{G}, \mathbb{G}_{\mathbb{T}}, \hat{e})$, where q is a k-bit prime number, $\mathbb{G}$ and $\mathbb{G}_{\mathbb{T}}$ are two groups with order q, $g \in \mathbb{G}$ is a generator, and $\hat{e} : \mathbb{G} \times \mathbb{G} \longrightarrow \mathbb{G}_{\mathbb{T}}$ is a nondegenerated and efficiently computable bilinear map.

C. 0-Encoding and 1-Encoding

According to [30], we introduce the 0-encoding and 1-encoding technique which will be used to identify the relationship between two vectors.

An integer s can be denoted as a binary vector $\vec{s} = (s_n, s_{n-1}, \ldots, s_1)$, where $s = \sum_{i=1}^{n} s_i \cdot 2^{i-1}$, and $s_i \in \{0, 1\}$.

The 0-encoding of $\vec{s}$ is the set S_s^0 of binary vectors such that

$$S_s^0 = \{(s_n, s_{n-1}, \ldots, s_{i+1}, 1) | s_i = 0, 1 \leq i \leq n\}. \quad (1)$$

The 1-encoding of $\vec{s}$ is the set S_s^1 of binary vectors such that

$$S_s^1 = \{(s_n, s_{n-1}, \ldots, s_i) | s_i = 1, 1 \leq i \leq n\}.$$

Both S_s^1 and S_s^0 have at most n elements.

Theorem 1: The integer $x > y$ if and only if S_x^1 and S_y^0 has a common element.

IV. Proposed CINEMA Framework

In this section, we first introduce a precise diagnosis model by using skyline computation, then we propose our CINEMA framework, as shown in Fig. 3, which consists of five phase: 1) system initialization; 2) data preparation; 3) query generation; 4) privacy-preserving online medical primary diagnosis service; and 5) query result reading. Specially, SP first provides registration for the user in the *system initialization* phase and executes some preprocess operation on skyline points extracted from large medical data set in the *data preparation* phase. Then the user preprocess the query vector which contains a lot of health information by introducing different random numbers in the *query generation* phase. After that, SP operates the preprocessed query vector with lightweight polynomial aggregation technique in the privacy-preserving online medical primary diagnosis service phase. Finally, the user obtains the final diagnosis result from SP in the *query result reading* phase. Meanwhile, for easier expression, we give the description of variables used in the following sections in Table I.

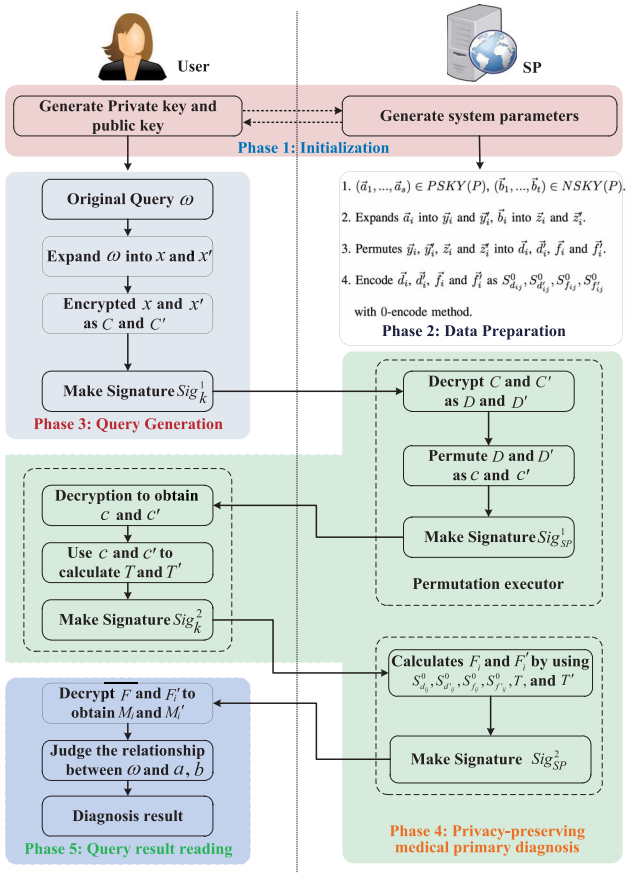


Fig. 3. Architecture of CINEMA.

TABLE I
Definition of Notations in CINEMA

Notation	Definition
$g, q, \mathbb{G}, \mathbb{G}_{\mathbb{T}}, \hat{e}$	The parameters of bilinear paring.
τ, q, η, ϕ	The secure parameters chosen by user.
$\vec{R}, \vec{R}'$	Random number vectors chosen by SP.
$\vec{v}_i, \vec{v}_i'$	Label vectors.
$\vec{w}$	The user's query vector.
$\vec{x}, \vec{x}'$	The expanded vectors of $\vec{w}$.
$k_i, R_i, c_{ei}, c_{\bar{ei}}$	Random numbers sequence.
$\vec{y}_i, \vec{y}_i', \vec{z}_i, \vec{z}_i'$	The expanded vectors.
$\vec{d}_i, \vec{d}_i', \vec{f}_i, \vec{f}_i'$	The permuted vectors.
$S_{d_{ij}}^0, S_{f_{ij}}^0$	0-encoding of d_{ij}, f_{ij}.
$T_{ei}, T_{\bar{ei}}, \vec{u}_i, \vec{u}_i'$	Binary vectors.
$\pi(\cdot), \pi'(\cdot)$	Permutation sequences.
$O_{d_{ij}(\beta)}, O_{d'_{ij}(\beta)}$	Binary string.
$\vec{r}_i, \vec{r}_i'$	Random number vectors chosen by user.
EQ_i	The encryption of query request.
$ERes_i$	The encryption of query response.
$H(\cdot)$	The secure cryptographic hash function.
$E(\cdot)$	The secure asymmetric encryption algorithm.

A. Skyline Diagnosis Model

Considering that a large medical dataset $P = \{P_1, \ldots, P_n\}$ in m-dimensional space, P_a and P_b are two different points in P.

Definition 3 (Positive Skyline): We define P_a positive dominated P_b, denoted by $P\text{dom}(P_a, P_b)$, if it satisfies the following conditions: 1) $\forall 1 \leq j \leq m$, $P_a[j] \leq P_b[j]$ and 2) at least there

exists one j, $P_a[j] < P_b[j]$, where $P_i[j]$ is the jth dimension of P_i and $1 \le i \le n$. The positive skyline set PSKY(P) contains lots of points which are not positive dominated by any other points in P, and the value of $P\text{dom}(P_a, P_b)$ can be defined as

$$P\text{dom}(P_a, P_b) = \begin{cases} 1, & \text{if } P_a \text{ positive dominated } P_b \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Definition 4 (Negative Skyline): We define P_a negative dominated P_b, denoted by $N\text{dom}(P_a, P_b)$, if it satisfies the following conditions: 1) $\forall 1 \le j \le m$, $P_a[j] \ge P_b[j]$ and 2) at least there exists one j, $P_a[j] > P_b[j]$, where $P_i[j]$ is the jth dimension of P_i and $1 \le i \le n$. The negative skyline set NSKY(P) contains lots of points which are not negative dominated by any other points in P, and the value of $N\text{dom}(P_a, P_b)$ can be defined as

$$N\text{dom}(P_a, P_b) = \begin{cases} -1, & \text{if } P_a \text{ negative dominated } P_b \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

Suppose that PSKY(P) $= \{a_1, \ldots, a_s\}$ and NSKY(P) $= \{b_1, \ldots, b_t\}$, $1 \le s, t \le n$, then the point a_i and b_j can present as vector $\vec{a}_i = (a_{i1}, \ldots, a_{im}) \in \mathbb{Z}_q^m$ and $\vec{b}_j = (b_{j1}, \ldots, b_{jm}) \in \mathbb{Z}_q^m$, where $1 \le i \le s$ and $1 \le j \le t$, and the query request can present as vector $\vec{w} = (w_1, \ldots, w_m) \in \mathbb{Z}_p^m$.

According to (2) and (3), we can define the skyline diagnosis standard as follows:

$$\begin{cases} \rho_+ = \frac{1}{s}\sum_{i=1}^{s} P\text{dom}(\vec{a}_i, \vec{w}) \\ \rho_- = \frac{1}{t}\sum_{j=1}^{t} N\text{dom}(\vec{b}_j, \vec{w}). \end{cases} \quad (4)$$

Only if $\rho_+ = 1$ and $\rho_- = -1$, we can confirm that the diagnosis result is positive.

B. System Initialization

SP first chooses a security parameter k and executes function Gen(k) to generate the bilinear parameters $(q, g, \mathbb{G}, \mathbb{G}_\mathbb{T}, \hat{e})$, Then, SP chooses a random number in $\mathbb{Z}_q^*$ as its private key SK_{SP}, and computes its public key $\text{PK}_{\text{SP}} = g^{\text{SK}_{\text{SP}}}$. In addition, SP chooses a secure asymmetric encryption algorithm $E()$, i.e., ElGamal, and a secure cryptographic hash function $H()$, where $H : \{0, 1\}^* \longrightarrow \mathbb{Z}_q^*$. After that, SP keeps its private key SK_{SP} as master key secretly, and publishes the system parameters $<q, g, \mathbb{G}, \mathbb{G}_\mathbb{T}, \hat{e}, \text{PK}_{\text{SP}}, E(), H()>$.

When registering in the SP, user U_K chooses a random number in $\mathbb{Z}_q^*$ as her/his private key SK_{U_K}, and computes its public key $\text{PK}_{U_K} = g^{\text{SK}_{U_K}}$, then submits her/his information and PK_{U_K} to SP through a secure channel for signature. SP first checks the correctness of the user's information. If it dose hold, SP makes a signature for PK_{U_K} with its private key SK_{SP} and send it back to U_K.

C. Data Preparation

Considering a precise medical dataset $P = \{P_1, .., P_n\}$, which contains lots of medical instances with m dimension. SP owns an skyline diagnosis model, which mainly consists of two parts: 1) positive skyline point set PSKY(P) and 2) negative skyline point set NSKY(P). Assume that PSKY(P) $= \{a_1, \ldots, a_s\}$ and NSKY(P) $= \{b_1, \ldots, b_t\}$, then each point $a_i \in Psky(P)$ and $b_j \in Nsky(P)$ can be expressed as vectors $\vec{a}_i = (a_{i1}, \ldots, a_{im}) \in \mathbb{Z}_q^m$ and $\vec{b}_j = (b_{j1}, \ldots, b_{jm}) \in \mathbb{Z}_q^m$, as the peripheral data of the medical dataset P, it is sensitive and need to be kept secret from the users.

C-1: For each vector $\vec{a}_i = (a_{i1}, \ldots, a_{im}) \in Psky(P)$, $1 \le i \le s$, SP first expands it into two $2m$-length vectors: $\vec{y}_i = (y_{i1}, \ldots, y_{i2m})$ and $\vec{y}'_i = (y'_{i1}, \ldots, y'_{i2m})$, where

$$\begin{cases} y_{ij} = 2a_{ij} \\ y_{i(n+j)} = -2a_{ij} \\ y'_{ij} = 2a_{ij} + 1 \\ y'_{i(n+j)} = -(2a_{ij} + 1) \end{cases} \quad (5)$$

$1 \le j \le m$. Then SP constructs two permutation sequences π, π', and two random number vectors $\vec{R}$, $\vec{R}'$, for each vector $\vec{y}_i$ and $\vec{y}'_i$, SP executes the following operations:

$$\begin{cases} \vec{d}_i = \pi(\vec{y}_i + \vec{R}) \\ \vec{d}'_i = \pi'(\vec{y}'_i + \vec{R}'). \end{cases} \quad (6)$$

Then, for each element $d_{ij} \in \vec{d}_i$, $d'_{ij} \in \vec{d}'_i$, where $1 \le j \le 2m$, SP generates the sets $S^0_{d_{ij}} = \{O_{d_{ij}(\beta)}, \ldots, O_{d_{ij}(1)}\}$ and $S^0_{d'_{ij}} = \{O_{d'_{ij}(\beta)}, \ldots, O_{d'_{ij}(1)}\}$ by using 0-encoding technique in (1).

C-2: For each vector $\vec{b}_i \in \text{NSKY}(P)$, $1 \le i \le t$, SP executes (5) on $\vec{b}_i$, $\vec{b}'_i$ to generate $\vec{z}_i$, $\vec{z}'_i$, then permutes $\vec{z}_i$, $\vec{z}'_i$ into $\vec{f}_i$, $\vec{f}'_i$ by executing (6), and encodes each element in $\vec{f}_i$ and $\vec{f}'_i$ as $S^0_{f_{ij}} = \{O_{f_{ij}(\beta)}, \ldots, O_{f_{ij}(1)}\}$, and $S^0_{f'_{ij}} = \{O_{f'_{ij}(\beta)}, \ldots, O_{f'_{ij}(1)}\}$ as well.

C-3: After that, SP constructs two $2m$-length label vector $\vec{v} = (1, \ldots, 1, 0, \ldots, 0)$ and $\vec{v}' = (1, \ldots, 1, 0, \ldots, 0)$, which are used to identify the relationship of each element in the vector. Specifically, if $v_i = 1$, then $x_i > y_i$ in position i. Otherwise, $x_i < y_i$. Similarly, if $v'_i = 1$, then $x' < y'$. Otherwise, $x'_i > y'_i$. Then, in order to keep $\vec{v}$ and $\vec{v}'$ secret from the users, SP computes $h = H(\pi(\vec{v}))$ and $h' = H(\pi'(\vec{v}'))$ with a cryptographic hash function $H()$. In the phase of query result reading, h and h' will be used as the criteria for judging the dominance relationship.

D. Query Generation

User U_K wants to request medical primary diagnosis service by sending a query vector $\vec{w} = (w_1, \ldots, w_m) \in \mathbb{Z}_q^m$, which contains lots of sensitive medical data, so $\vec{w}$ needs to be kept secret from the trusted-and-curious SP during the process of query.

D-1: U_K first expands the query vector $\vec{w}$ into two $2m$-length vectors: $\vec{x} = (x_1, \ldots, x_{2m})$ and $\vec{x}' = (x'_1, \ldots, x'_{2m})$, where

$$\begin{cases} x_i = 2w_i + 1 \\ x_{m+i} = -(2w_i + 1) \\ x'_i = 2w_i \\ x'_{m+i} = -2w_i \end{cases}$$

and $1 \le i \le m$.

D-2: U_K chooses three large prime numbers τ, q, and η, and $2m$ large positive random numbers $r_1, \ldots, r_{2m}$, where $|\eta| < |q|$. Considering $C_0 = \tau^{-1} \bmod q$, then U_K calculates $p = C_0 + k_0 \cdot q$ and $\phi = p \cdot \eta$. For each $j \in (1, 2, \ldots, 2m)$, U_K executes the following operations:

$$\begin{cases} C_j = \phi \cdot r_j + x_j \bmod q \\ C'_j = \phi \cdot r_j + x'_j \bmod q. \end{cases} \quad (7)$$

Let $Q_1 = <\phi \| q \| C_1 \| \ldots \| C_{2m} \| C'_1 \| \ldots \| C'_{2m} \| TS_1>$, TS_1 is the current timestamp, which is used to resist the potential relay attack. U_K keep τ, η and p secret, and creates signature $\mathrm{Sig}_K^1 = (H(Q_1))^{\mathrm{SK}_{U_K}}$ by private key SK_{U_K}, then computes the query request $EQ_1 = E_{\mathrm{PK}_{\mathrm{SP}}}(Q_1 || \mathrm{Sig}_K^1)$. Finally, U_K sends EQ_1 to SP.

E. Privacy-Preserving Diagnosis and Response

After receiving EQ_1, SP decrypts EQ_1 with its secret key $\mathrm{SK}_{\mathrm{SP}}$ to obtain Q_1 and Sig_K^1 first, then verifies its validity by checking whether $\hat{e}(g, \mathrm{Sig}_K^1) = \hat{e}(\mathrm{PK}_{U_k}, H(Q_1))$. If it dose hold, the packet is valid. Then SP obtains $\phi, q, C_1, \ldots, C_{2m}, C'_1, \ldots, C'_{2m}$.

E-1: For each $j \in (1, 2, \ldots, 2m)$, SP chooses $2m$ random numbers $k_1, \ldots, k_{2m}$ to compute

$$\begin{aligned} D_j &= C_j + \phi \cdot k_j + R_j \bmod q \\ &= \phi \cdot k_j + (C_j + R_j) \bmod q \end{aligned} \tag{8}$$

where $R_j \in \vec{R}$, then SP obtains the vector $\vec{D} = (D_1, \ldots, D_{2m})$. By executing the similar operation with (8) on C'_j, SP obtains the vector $\vec{D}' = (D'_1, \ldots, D'_{2m})$. After that, SP uses π and π' to execute permutation operation on vector $\vec{D}$ and $\vec{D}'$ to generate $\pi(\vec{D})$ and $\pi'(\vec{D}')$.

E-2: Let $\mathrm{Res}_1 = <\pi(\vec{D}) || \pi'(\vec{D}') || TS_2>$, TS_2 is the current timestamp, which is used to resist the potential relay attack. SP creates signature $\mathrm{Sig}_{\mathrm{SP}}^1 = (H(\mathrm{Res}_1))^{\mathrm{SK}_{\mathrm{SP}}}$ by private key $\mathrm{SK}_{\mathrm{SP}}$, then computes the response $E\mathrm{Res}_1 = E_{\mathrm{SK}_{\mathrm{SP}}}(\mathrm{Res}_1 || \mathrm{Sig}_{\mathrm{SP}}^1)$. Finally, SP sends $E\mathrm{Res}_1$ to U_K.

E-3: After receiving the $E\mathrm{Res}_1$, U_K decrypts $E\mathrm{Res}_1$ with its secret key SK_{U_K} to obtain R_1 and $\mathrm{Sig}_{\mathrm{SP}}^1$ first, then verifies its validity by checking whether $\hat{e}(g, \mathrm{Sig}_{\mathrm{SP}}^1) = \hat{e}(\mathrm{PK}_{\mathrm{SP}}, H(\mathrm{Res}_1))$. If it dose hold, the packet is valid, then U_K obtain $\pi(D_1), \ldots, \pi(D_{2m})$ and $\pi'(D'_1), \ldots, \pi'(D'_{2m})$, and execute the following operations:

$$\begin{aligned} E_j &= p^{-1} \cdot D_j \bmod q \\ &= \eta \cdot k_j + \pi(C_j + R_j) \cdot p^{-1} \end{aligned} \tag{9}$$

$$\begin{aligned} \frac{E_j \bmod \eta}{\tau} &= \frac{\eta \cdot k_j + \pi(C_j + R_j) \cdot p^{-1} \bmod \eta}{\tau} \\ &= \pi(x_j + R_j) \end{aligned} \tag{10}$$

where $1 \le j \le 2m$. Denote that $\vec{c} = \pi(\vec{x} + \vec{R}) = (c_1, \ldots, c_{2m})$. Similarly, by executing the same operation as (5) and (6), U_k obtains the permuted vector $\vec{c}\,'$.

E-4: For each $c_j \in \vec{c}$, U_K converts c_j into binary vector $\vec{e}_j = (e_{j1}, \ldots, e_{jl})$, and chooses $2l$ random numbers $r_{e1}, \ldots, r_{el}, r_{\bar{e}1}, \ldots, r_{\bar{e}l}$, for each $i \in (1, \ldots, l)$, U_K executes the following operations:

$$\begin{cases} T_{ei} = \phi \cdot r_{ei} \bmod q \\ T_{\bar{e}i} = \phi \cdot r_{\bar{e}i} + 1 \bmod q. \end{cases} \tag{11}$$

According to (7), it is obvious that T_{ei} is equivalent to the encryption of 1, and $T_{\bar{e}i}$ is equivalent to the encryption of 0. Then U_K obtain the vector $T = (T_{e1}, \ldots, T_{el}, T_{\bar{e}1}, \ldots, T_{\bar{e}l})$. By executing the similarly operations as (11), U_K obtains the vector $T' = (T'_{e1}, \ldots, T'_{el}, T'_{\bar{e}1}, \ldots, T'_{\bar{e}l})$.

E-5: Let $Q_2 = <T_{e1} \| \ldots \| T_{\bar{e}l} \| T'_{el} \| \ldots \| T'_{\bar{e}l} \| TS_3>$, TS_3 is the current timestamp, which is used to resist the potential relay attack. U_K creates signature $\mathrm{Sig}_K^2 = (H(Q_2))^{\mathrm{SK}_{U_K}}$ by private key SK_{U_K}, then computes the query request $EQ_2 = E_{\mathrm{PK}_{\mathrm{SP}}}(Q_2 || \mathrm{Sig}_K^2)$. Finally, U_K sends EQ_2 to SP.

E-6: After receiving EQ_2, SP decrypts EQ_2 with its secret key $\mathrm{SK}_{\mathrm{SP}}$ to obtain Q_2 and Sig_K^2 first, then verifies its validity by checking whether $\hat{e}(g, \mathrm{Sig}_K^2) = \hat{e}(\mathrm{PK}_{U_k}, H(Q_2))$. If it dose hold, the packet is valid, then U_K obtain $T = (T_{e1}, \ldots, T_{el}, T_{\bar{e}1}, \ldots, T_{\bar{e}l})$ and $T' = (T'_{e1}, \ldots, T'_{el}, T'_{\bar{e}1}, \ldots, T'_{\bar{e}l})$.

E-7: For each binary vector $O_{d_{ij}(\beta)} = (\theta_1, \ldots, \theta_\gamma) \in S^0_{d_{ij}}$ and $O'_{d_{ij}(\beta)} = (\theta'_1, \ldots, \theta'_\gamma) \in S^0_{d'_{ij}}$, where the values of θ_i and θ'_i are either 1 or 0, $1 \le \theta \le \gamma$. Assume the number of elements in $S^0_{d_{ij}}$ and $S^0_{d'_{ij}}$ are δ and δ', by using the random number vector $\vec{R}'$, SP calculates

$$\begin{cases} G_i = T_{\theta_\gamma} + \cdots + T_{\theta_1} + \phi \cdot R'_i \bmod q \\ G_i{}' = T_{\theta_{\gamma'}} + \cdots + T_{\theta_1'} + \phi \cdot R'_i \bmod q \end{cases} \tag{12}$$

where $1 \le i \le \delta$, then SP chooses two random number vectors $\vec{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_{l-\delta})$ and $\vec{\varepsilon}\,' = (\varepsilon'_1, \ldots, \varepsilon'_{(l-\delta')})$, and permutes $\vec{G}_i$ with $\vec{\varepsilon}$, and $\vec{G}'_i$ with $\vec{\varepsilon}\,'$, then denotes the permuted results as $\vec{F}_a = (F_{a1}, \ldots, F_{al})$ and $\vec{F}'_a = (F'_{a1}, \ldots, F'_{al})$. By executing the similarly operations as (12), SP obtains the vector $\vec{F}_b = (F_{b1}, \ldots, F_{bl})$ and $\vec{F}'_b = (F'_{b1}, \ldots, F'_{bl})$.

E-8: Let $\mathrm{Res}_2 = <\vec{F}_a || \vec{F}'_a || \vec{F}_b || \vec{F}'_b || TS_4>$, TS_4 is the current timestamp, which is used to resist the potential relay attack. SP creates signature $\mathrm{Sig}_{\mathrm{SP}}^2 = (H(\mathrm{Res}_2))^{\mathrm{SK}_{\mathrm{SP}}}$ by private key $\mathrm{SK}_{\mathrm{SP}}$, then computes the response $ER_2 = E_{\mathrm{SK}_{\mathrm{SP}}}(\mathrm{Res}_2 || \mathrm{Sig}_{\mathrm{SP}}^2)$. Finally, SP sends $E\mathrm{Res}_2$ to U_K.

F. Query Result Reading

After receiving the $E\mathrm{Res}_2$, U_K decrypts $E\mathrm{Res}_2$ with its secret key SK_{U_K} to obtain Res_2 and $\mathrm{Sig}_{\mathrm{SP}}^2$ first, then verifies its validity by checking whether $\hat{e}(g, \mathrm{Sig}_{\mathrm{SP}}^2) = \hat{e}(\mathrm{PK}_{\mathrm{SP}}, H(\mathrm{Res}_2))$. If it dose hold, the packet is valid, then U_K obtain $\vec{F}_a$, $\vec{F}'_a$, $\vec{F}_b$, $\vec{F}'_b$.

F-1: For each $\xi \in (1, \ldots, l)$, U_K executes the following operations:

$$\begin{cases} N_{a\xi} = p^{-1} \cdot F_{a\xi} \bmod q \\ N'_{a\xi} = p^{-1} \cdot F'_{a\xi} \bmod q \end{cases} \tag{13}$$

and

$$\begin{cases} M_{a\xi} = \frac{N_{a\xi} \bmod \eta}{\tau} \\ M'_{a\xi} = \frac{N'_{a\xi} \bmod \eta}{\tau}. \end{cases} \tag{14}$$

If there exists some $1 \le \xi \le l$ such that $M_{a\xi} = 0$, then U_k knows that $c_j > d_{ij}$, and denotes that as

$$u_{ij}^{(a)} = \begin{cases} 1, & \text{only if } c_j > d_{ij} \\ 0, & \text{otherwise.} \end{cases}$$

Similarly, if there exists some $1 \le \xi \le l$ such that $M'_{a\xi} = 0$, then U_k knows that $c'_j > d'_{ij}$, and denotes that as

$$u'^{(a)}_{ij} = \begin{cases} 1, & \text{only if } c'_j < d'_{ij} \\ 0, & \text{otherwise} \end{cases}$$

where $1 \le i \le s$, $1 \le j \le 2m$, then U_K obtains the vector $\vec{u}_{ai} = (u_{i1}^{(a)}, \ldots, u_{im}^{(a)})$ and $\vec{u}'_{ai} = (u'^{(a)}_{i1}, \ldots, u'^{(a)}_{im})$. Similarly, U_K can also obtains vectors $\vec{u}_{bi}$ and $\vec{u}'_{bi}$.

F-2: U_k calculates $H(\vec{u}_{ai})$ and $H(\vec{u}'_{ai})$ with the hash function $H()$, then verifies whether $H(\vec{u}_{ai}) = h$ and $H(\vec{u}'_{ai}) = h'$ hold. As a result, there are three cases below.

Case 1: If $H(\vec{u}_{ai}) = h$, $P\text{dom}(\vec{w}, \vec{a}_i) = 1$.

Case 2: If $H(\vec{u}'_{ai}) = h'$, $N\text{dom}(\vec{w}, \vec{a}_i) = -1$.

Case 3: If $H(\vec{u}_{ai}) \neq h$ and $H(\vec{u}'_{ai}) \neq h'$, $\vec{w}$ and $\vec{a}_i$ have no dominance relationship, and $P\text{dom}(\vec{w}, \vec{a}_i) = 0$.

According to above, U_k can also obtains the relationship between vector $\vec{w}$ and $\vec{b}_i$.

F-3: According to (3), for each $i \in (1, s)$ and $j \in (1, t)$, U_K calculates

$$\begin{cases} \rho_+ = \frac{1}{s}\sum_{i=1}^{s} P\text{dom}(\vec{w}, \vec{a}_i) \\ \rho_- = \frac{1}{t}\sum_{j=1}^{t} N\text{dom}(\vec{w}, \vec{b}_j). \end{cases}$$

Only if $\rho_+ = 1$ and $\rho_- = -1$, then U_k knows she/he is more likely suffers from the disease.

Correctness: In (9) and (10), in order to ensure the correctness of the decryption, considering the aforementioned constraints, ie,. $|\eta r + xp^{-1}| < |q|$, $|xp^{-1}| < |\eta|$, and $|q| < |\phi|$, such that p is prime, while in (14), to correctly decrypt the result, the following constraints should be satisfied: $|\eta \sum_{j=i}^{s} c_{t_j} + R + sp^{-1}| < |q|, |sp^{-1}| < |\eta|, |q| < |\phi|$. We can just set $|\eta| = k_3 = 512$, $|q| = k_q = 1024$, $|\tau| = 256$, $|r| = 512$, $|p| = k_2 = 1536$, and $|\phi| = 2048$.

V. Security Analysis

In this section, we analyze the security properties of the proposed CINEMA framework. Especially, according to the security requirements discussed in Section II, our security analysis will mainly focus on two parts: 1) how the CINEMA framework protects the privacy of the user's query request and 2) ensures the confidentiality of skyline diagnosis model simultaneously.

A. Privacy of User's Query Request

During the query generation phase, the query vector $\vec{w}$ is expanded into two $2m$-length vectors $\vec{x}$ and $\vec{x}'$ first, for each number x_i and x'_i, user calculates $C_j = \phi \cdot r_j + x_j \bmod q$ and $C'_j = \phi \cdot r_j + x'_j \bmod q$, where $C_0 = \tau^{-1} \bmod q$, $p = C_0 + k_0 \cdot q$, and sends the vectors $\vec{C} = (C_1, \ldots, C_{2m})$ and $\vec{C}' = (C'_1, \ldots, C'_{2m})$ to SP. Since the random number r_j is individually used for once, different r_j are indistinguishable, and τ, η, and p are kept secret by registered users, SP cannot obtain any sensitive medical data from the vectors $\vec{C}$ and $\vec{C}'$.

During the privacy-preserving diagnosis and response phase, when the user obtains the permuted vector $\vec{c} = \pi(\vec{x} + \vec{R})$, then converts each number $c_j \in \vec{c}$ into l-length binary vector $\vec{e}_j = (e_{j1}, \ldots, c_{jl})$, and chooses $2l$ random numbers $r_{e1}, \ldots, r_{el}, r_{\bar{e}1}, \ldots, r_{\bar{e}l}$, then computes $T_{ei} = \phi \cdot r_{ei} \bmod q$ (which is equivalent to the encryption of 0), and $T_{\bar{e}i} = \phi \cdot r_{\bar{e}i} + 1 \bmod q$ (which is equivalent to the encryption of 1). Since the random numbers r_{ei} and $r_{\bar{e}i}$ are individually used for once, different r_{ei} and $r_{\bar{e}i}$ are indistinguishable, SP can only obtain the vector which consist of 0 and 1, without any information about the vector $\vec{x}$. Similarly, SP can not obtains any information about the vector $\vec{x}'$.

Moreover, the user's encrypted query is transmitted under secure channel, and only SP can obtain the encrypted query vector. Thus, the user's query information is privacy-preserving during the computation.

B. Confidentiality of Diagnosis Model

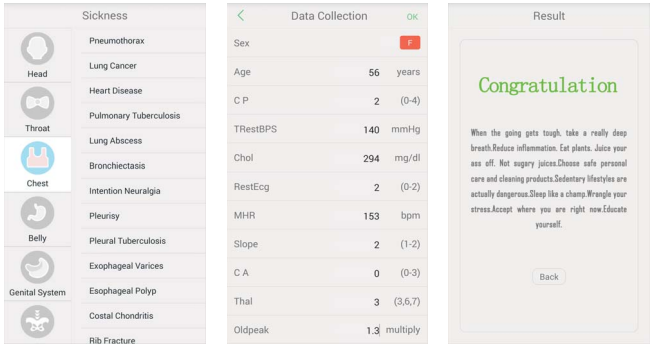
In our proposed CINEMA framework, SP owns a skyline diagnosis model which consist of medical dataset PSKY(P) and NSKY(P), during the phase of data preparation, SP generates the permutation sequences π and π', random number vectors $\vec{R}$ and $\vec{R}'$. For SP, the data in PSKY(P) and NSKY(P) is the intellectual property that should be protected, as well as the permutation sequences π, π', and random number vectors $\vec{R}$, $\vec{R}'$, which are the security parameters of the framework.

In the phase of data preparation, for each vector $\vec{a}_i = (a_{i1}, \ldots, a_{im}) \in$ PSKY(P), SP first expands them into two $2m$-length vectors $\vec{y}_i$ and $\vec{y}'_i$, where $1 \leq i \leq s$, then calculates $\vec{d}_i = \pi(\vec{y}_i + \vec{R})$, $\vec{d}'_i = \pi'(\vec{y}'_i + \vec{R}')$ with permutation sequences π, π', and random number vector $\vec{R}$, $\vec{R}'$. After that, SP constructs two $2m$-length label vectors $\vec{v}$ and $\vec{v}'$, in order to prevent the user from knowing π and π' by comparing the elements in $\vec{v}$ and $\pi(\vec{v})$, $\vec{v}'$ and $\pi(\vec{v})'$, SP uses hash function $H()$ to calculate $h = H(\pi(\vec{v}))$ and $h' = H(\pi'(\vec{v}'))$.

During the phase of privacy-preserving medical primary diagnosis service, when obtains the vectors $\vec{C}$ and $\vec{C}'$ from the user, for each element $C_j \in \vec{C}$ and $C'_j \in \vec{C}'$, SP chooses two random numbers k_j and k'_j to computes $D_j = C_j + \phi \cdot k_j + R_j \bmod q$, $D'_j = C'_j + \phi \cdot k'_j + R'_j \bmod q$, where $R_j \in \vec{R}$ and $R'_j \in \vec{R}'$, then permutes D_j and D'_j with permutation sequence π and π'. SP sends $\pi(\vec{D})$ and $\pi'(\vec{D}')$ to the user. When the user obtains $\pi(x_j + R_j)$ and $\pi'(x'_j + R'_j)$ by decrypting D_j and D'_j, since π and π' are kept secret by SP, the user can not obtain the real value of the random number R_j, R'_j by comparing $\pi(x_j + R_j)$ and x_j, $\pi'(x'_j + R'_j)$ and x'_j. Thus, the vectors $\vec{R}$ and $\vec{R}'$ are security.

Next, we will analyze how the proposed CINEMA framework can protect the privacy of PSKY(P) and NSKY(P) during the phase of privacy-preserving medical diagnosis service. For each binary vector $O_{d_{ij(\beta)}} \in S^0_{d_{ij}}$ and $O_{d'_{ij(\beta)}} \in S^0_{d'_{ij}}$, SP hide the selection of T_{θ_i} and $T_{\theta'_i}$ by calculating $G_i = T_{\theta_\gamma} + \cdots + T_{\theta_1} + \phi \cdot R'_i \bmod q$ (which is equivalent to the encryption of $(T_{\theta_\gamma} + \cdots + T_{\theta_1})$) and $G'_i = T_{\theta'_\gamma} + \cdots + T_{\theta'_1} + \phi \cdot R'_i \bmod q$ [which is equivalent to the encryption of $(T_{\theta'_\gamma} + \cdots + T_{\theta'_1})$]. After that, SP chooses two random number vectors $\vec{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_{l-\delta})$ and $\vec{\varepsilon}\,' = (\varepsilon'_1, \ldots, \varepsilon'_{(l-\delta')})$, then permutes $G_1, \ldots, G_\delta$ with $\varepsilon_1, \ldots, \varepsilon_{l-\delta}$, $G'_1, \ldots, G'_\delta$ with $\varepsilon'_1, \ldots, \varepsilon'_{(l-\delta')}$, and denotes the permuted results as $\vec{F} = (F_1, \ldots, F_l)$ and $\vec{F}' = (F'_1, \ldots, F'_l)$. As a result, the user cannot know exactly how many elements contained in $S^0_{d_{ij}}$ and $S^0_{d'_{ij}}$. when the user obtains the vector $\pi(\vec{G}, \vec{\varepsilon})$ and $\pi'(\vec{G}', \vec{\varepsilon}\,')$ by decrypting $\vec{F}$ and $\vec{F}'$. Since the random numbers ε_i, ε'_i, and permutation sequence π, π' are keep secret by SP, the user can only knows the relationship between the vectors $\vec{w}$ and $\vec{a}_i$, $\vec{w}$ and $\vec{b}_i$, while gains nothing else about the data in PSKY(P) and NSKY(P).

Moreover, the response from SP is transmitted under secure channel, and only the valid user can obtain the encrypted



(a) (b) (c)

Fig. 4. Implementation of CINEMA. (a) Select the type of disease. (b) Input medical data. (c) Prediagnosis result.

response. Thus, the confidentiality of the skyline diagnosis model can be achieved during the computation.

C. Authentication of Query Request

In the proposed CINEMA framework, each registered user's request is signed by Boneh–Lynn–Shacham (BLS) short signature [31]. Since the BLS short signature is provably secure under the oracle model, the source authentication was guaranteed. Moreover, any unregistered user cannot submit valid query request to SP without the valid secret key, she/he also cannot submit valid query request to SP. As a result, the query request from the unregistered user can be detected in the proposed CINEMA framework.

From the above analysis, we can conclude that the proposed CINEMA framework is secure and privacy-preserving both for user and SP, and the security requirements are satisfied as well.

VI. Performance Evaluations

In this section, we first evaluate the accuracy and computational complexity of the proposed CINEMA framework. Then, we implement CINEMA framework and deploy it in the real environment to evaluate its integrated performance.

A. Evaluation Environment

In order to measure the comprehensive performance of CINEMA in the real environment, we implement CINEMA on computers with a real medical database. Specifically, two computers with 2.3 GHz six-core processor, 16 GB RAM, and Windows 7 operating system are chosen to evaluate the client and SP, respectively, which are connected through 802.11g WLAN. Based on CINEMA framework, we construct two simulators in Java and run the implementation on computers to simulate the client and SP. As shown in Fig. 4, a user chooses the type of disease and entering some physiological data through the client, then get the diagnosis result from SP. In particular, in order to obtain the correct diagnosis result, we can just set $|\eta| = k_3 = 512$, $|q| = k_1 = 1024$, $|\tau| = 256$, $|r| = 512$, $|p| = k_2 = 1536$, and $|\phi| = 2048$. In addition, we consider one real dataset to evaluate the accuracy of our proposed framework. It is from the UCI machine learning repository called Heart Disease Data (HDD) set [32].

TABLE II
Comparison of Accuracy

Accuracy	CINEMA	CPLD
Yes(495)	486(98%)	486(98%)
No(404)	387(95%)	387(95%)
Overall(899)	873(97%)	873(97%)

TABLE III
Comparison of Computation Complexity

	CINEMA	CDSS [17]
User	$2n \cdot T_{hash} + (16m + 8ml) \cdot T_{mul}$	$(5m + 4n + 4) \cdot T_{exp} + (6m + 3n + 5) \cdot T_{mul}$
SP	$4m(\sigma \cdot n + n + 2) \cdot T_{mul}$	$(mn + 7n + 9m + 9) \cdot T_{exp} + (5mn + 11n + 12m + 9) \cdot T_{mul}$

B. Accuracy Evaluation

According to the HDD dataset, we randomly choose 899 instances from cleveland.data, hungarin.data, switzerland.data. Each instance contains 75 attributes, we extract the main 12 attributes that may closely related to the heart disease, such as age in years, chest pain type, resting BP in mm/Hg, serum cholesterol in mg/dl, fasting blood sugar, resting electrocardiographic results, maximum heart rate, exercise-induced angina, old peak, the slope of the peak exercise ST segment, the number of major vessels colored by fluoroscopy, the year of cardiac cath. Before training the skyline diagnosis model, all the instances should be normalized. Then, we use the skyline diagnosis model and HDD dataset to test the success rate in the plain domain (abbreviated as CLPD). Meanwhile, we also use the same skyline diagnosis model to evaluate the accuracy of our proposed CINEMA framework with the same simulation environment. Thus, we obtained Table II. From the table, we can see that the total number of correctly diagnosed heart disease instances is 486 out of 495 and that of nonheart disease instances is 387 out of 404. In total, 873 samples are correctly classified out of 899(97%), our privacy-preserving algorithm does not compromise the accuracy, and the test result also confirms it by achieving the same accuracy as that of CLPD.

C. Computation Complexity

The proposed CINEMA framework can provide efficient medical diagnosis service to medical users. Especially, we assume that there are n vectors in the skyline diagnosis model (which contains s positive skyline points and t negative skyline points, $n = s + t$), each vector is m-dimensional, as well as the query vector.

During the phase of query generation, the user first expands the query vector $\vec{w}$ into two $2m$-length vectors $\vec{x}$ and $\vec{x}'$, which requires $8m$ multiplication operations. when to encrypt the each element in $\vec{x}$ and $\vec{x}'$ by computing $C_j = \phi \cdot r_j + x_j \bmod q$ and $C'_j = \phi \cdot r_j + x'_j \bmod q$, the user have to execute $4m$ multiplication operations. After receiving the ciphertext from the user, SP cost $8m$ multiplication operations to generate $\vec{D} = (D_1, \ldots, D_{2m})$, $\vec{D}' = (D'_1, \ldots, D'_{2m})$ and permutes them as $\pi(\vec{D})$, $\pi'(\vec{D}')$. After receiving the response from SP, in order to obtain the permuted vectors $\vec{c} = \pi(\vec{x} + \vec{R})$ and

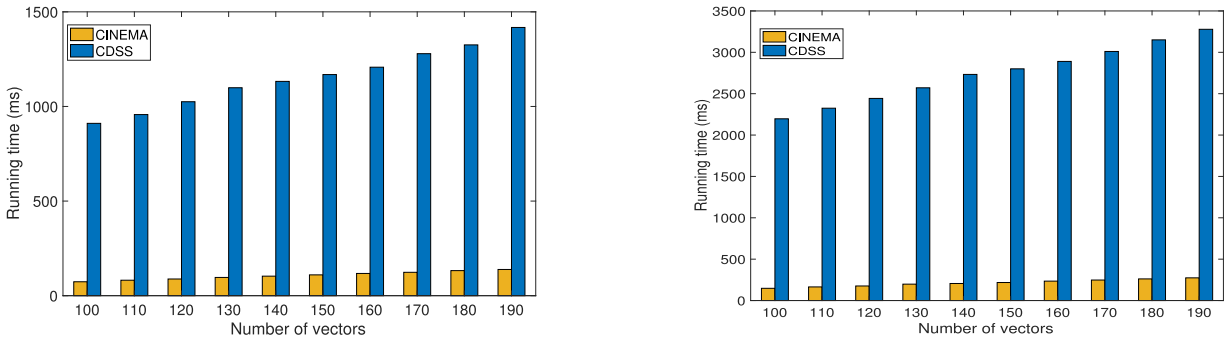


Fig. 5. Average running time in CINEMA versus CDSS. Average running time of (a) SP and (b) user.

$\vec{c}\,' = \pi'(\vec{x}'+\vec{R}')$, the user costs $4m$ multiplication operations to decrypt $\pi(\vec{D})$ and $\pi'(\vec{D}')$. After that, the user will cost $4m \cdot 2l$ multiplication operations to convert each number $c_j \in \vec{c}$ and $c'_j \in \vec{c}\,'$ into binary vectors and generates the vectors T and T', where l is the length of the binary vector. After receiving the vectors T and T', in order to compare the dominance relationship between $\vec{w}$ and $\vec{a}$, $\vec{w}$ and $\vec{b}$, SP first calculates the $(G_1, \ldots, G_\sigma)$ and $(G'_1, \ldots, G'_\sigma)$, then permutes them with random numbers $(\varepsilon_1, \ldots, \varepsilon_{l-\delta})$ and $(\varepsilon'_1, \ldots, \varepsilon'_{(l-\delta')})$ to generate $\vec{F}_i$ and $\vec{F}_i{}'$, which would cost $(4m+4m\cdot\varepsilon)n$ multiplication operations. After receiving the vectors $\vec{u}$ and $\vec{u}'$, user calculates $H(\vec{u})$ and $H(\vec{u}')$ by using hash function, it will cost $2n \cdot T_{\text{hash}}$ multiplication operations. As a result, the total computation cost for the user and SP is $(16m + 8ml) \cdot T_{\text{mul}} + 2n \cdot T_{\text{hash}}$ and $4m(\sigma \cdot n + n + 2)T_{\text{mul}}$, where T_{mul} is the running time of the multiplication and T_{hash} is the running time of the hash function.

Different from fully and partially homomorphic encryption techniques, which have a heavy computation and communication cost. By using fast secure permutation and comparison techniques, the proposed CINEMA can achieve high accuracy while preserving the privacy of users' medical data and the confidentially of skyline diagnosis model. Meanwhile, the computation is efficient. In order to compare with CINEMA, we present a clinical decision support system named CDSS [26], which preserves the privacy of user's medical data with Paillier encryption technique during the process of the query. We assume the dimension of query vector is m and the number of vectors in the cloud is n, as shown in Table III, the corresponding computational costs of the user and SP are $(5m + 4n + 4) \cdot T_{\text{exp}} + (6m + 3n + 5) \cdot T_{\text{mul}}$ and $(mn + 7n + 9m + 9) \cdot T_{\text{exp}} + (5mn + 11n + 12m + 9) \cdot T_{\text{mul}}$.

We present the computation complexity comparison of CINEMA and CDSS in Table I. It is obvious that our proposed CINEMA framework can achieve efficient medical diagnosis with low computation complexity for users and SP. To further demonstrate the advantage of CINEMA compared with CDSS, we evaluate the computation overhead in the environment described in Section VI-A. Fig. 5(a) and (b) depicts the computation overhead varying with the dimension of the query vector and the number of vectors in SP and the user, and we assume the dimension of each vector is 10. Through comparing Figs. 5 and 6, we can find that with the increase of the numbers of vectors, the computation overhead of CDSS significantly increases and it is much higher than that of our proposed CINEMA framework. Although the computation overhead of our proposed CINEMA framework also increases when the number of vectors is large, it is still much lower than that of CDSS. In conclusion, our proposed CINEMA framework can achieve better efficiency on computation overhead in user and SP.

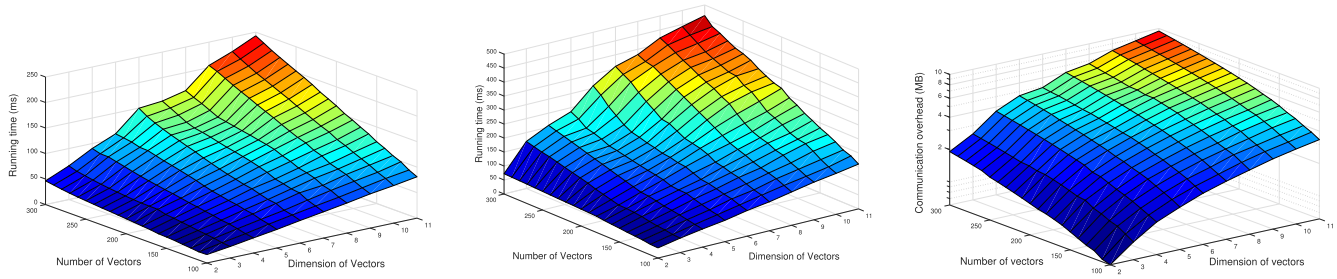
D. Efficiency Evaluation

In order to test the factors that may affect the efficiency of our proposed CINEMA, different skyline diagnosis model are randomly generated. We evaluate the computation cost and communication overhead in SP and user, respectively.

1) Computation Cost: We first evaluate the main factors that impact the computation cost of SP and the user.

a) SP: According to the definition of the skyline computation, we can note that the dimensions of vectors and the numbers of vectors may be the main factors that impact the computation complexity on SP in our proposed CINEMA framework. Therefore, different dimensions and numbers of vectors are chosen to illustrate the computation cost. The dimensions of vectors are selected from 2 to 11, and the numbers of vectors are chosen from 100 to 300. In order to ensure the accuracy, we execute the experiment 1000 times with different dimensions, as well as different numbers. As shown in Fig. 6(a), we can learn that the computation overhead of SP is increased with the increase of the vectors' dimension and number. The reason is that, when SP intends to offer medical diagnosis service for users, the vectors in SP will be permuted and operated to generate $\vec{F}_a$, $\vec{F}'_a$, $\vec{F}_b$, $\vec{F}'_b$, which cost much more time with the increase of vectors' dimension and number. However, due to the fact that basic operations are based on fast permutation and comparison techniques, the maximum time required for SP is less than 250 ms under the predefined evaluation environment.

b) User: The query response time of user (i.e., computer) is an important metric to illustrate the feasibility of our proposed CINEMA framework. Due to the dimensions of the



(a) (b) (c)

Fig. 6. Performance evaluation of CINEMA. Computation cost of (a) SP and (b) user. (c) Communication cost.

query vector is equal to the vectors in SP, therefore, different dimensions of the query vector are chosen to illustrate the computation cost of the computer. To observe the computation cost, different dimensions of the query vector are selected from 2 to 11, and the numbers of vectors in SP are chosen from 100 to 300. Then, we execute the experiment 1000 times with different dimensions and numbers. Specifically, Fig. 6(b) shows the computation overhead of computer is increased with the increasing of vectors' dimensions and numbers. The reason is that, the computer expands the query vector first, then encrypts the expanded results and send it to SP. Upon receiving the diagnosis result from the SP, the computer decrypts it and calculates the final diagnosis result, which cost much more time with the increasing of the vectors' dimensions and numbers. Similarly, due to the fact that basic operations are based on fast permutation and comparison techniques, the maximum time required for the computer is less than 500 ms under the predefined evaluation environment.

2) Communication Overhead: In CINEMA framework, the user's query packet is $EQ_1 = E_{\mathrm{PK_{SP}}}(Q_1||\mathrm{Sig}_K^1)$, and the first response form the SP is $E\mathrm{Res}_1 = E_{\mathrm{SK_{SP}}}(\mathrm{Res}_1||\mathrm{Sig}_{\mathrm{SP}}^1)$, then the user sends another packet $EQ_2 = E_{\mathrm{PK_{SP}}}(Q_2||\mathrm{Sig}_K^2)$ to SP, and finally get the diagnosis result $E\mathrm{Res}_2 = E_{\mathrm{SK_{SP}}}(\mathrm{Res}_2||\mathrm{Sig}_{\mathrm{SP}}^2)$. In the real environment, we record the size of the packets, as shown in Fig. 6(c), with the increase of the dimensions and numbers of vectors in SP, the communication overhead of CINEMA increases as well, when the dimension of the vectors is 11 and the number of vectors is up to 300, the total communication cost of the SP and the user is less than 10 MB under the predefined evaluation environment.

From the aforementioned analysis, we can conclude that our proposed CINEMA framework is indeed efficient in terms of computation and communication cost, which is suitable for providing online medical primary diagnosis service.

VII. Related Work

In this section, we will introduce some related works on skyline computation and privacy-preserving technique.

A. Skyline Computation

The skyline operator was first formalized by Borzsony *et al.* [6] with algorithm called block nested loop and divide and conquer. Thereafter, it was widely studied for building user's personalized queries over multidimensional datasets. Several sequential skyline algorithms [7]–[9] have been designed on efficiency for centralized storage, and the *Z*-search algorithm proposed by Mingjie *et al.* [9] was the state-of-the-art skyline computation algorithm. Recently, skyline computation for distributed database has received more attention. Liu *et al.* [11] proposed a skyline computation framework across multiple domains, within the framework, a skyline result from multiple SPs will be securely computed to provide better services for the client. Park *et al.* [12] constructed a quadtree for sampling data and judging the dominance relationships among different partitions, while the effect was not perfect. Mullesgaard *et al.* [33] represented grid-based partitioning by using a bit-string, which enables pruning more data points before final skyline computation. However, both centralized skyline and distributed skyline computation were well studied on improving the efficiency, while little of the works considered on similarity search. Kossmann *et al.* [34] proposed nearest neighbor algorithm which used the existing nearest neighbor search to split the data space recursively, while the privacy issue was overlooked. By exploring a novel neighboring relationship among POIs, Chen *et al.* [13] proposed several schemes that enable efficient verification of any location-based skyline query's result returned by an untrusted SP. In order to select the similar (or best) medical record over encrypted database, Liu *et al.* [14] proposed a fully secure skyline query protocol on data encrypted using semantically secure encryption, while the overhead of computation is heavy. Moreover, Lu *et al.* [35] pointed out that the conventional query over an encrypted database was not suitable for big data processing. Therefore, the more efficient secure skyline computation framework should be redesigned to fit for big data environment.

B. Privacy-Preserving Technique

Traditional anonymization techniques such as *k*-anonymity [20] and *l*-diversity [21], which through removes the personal identifiers (such as name and SSN) and obfuscating the quasi-identifiers (such as age, zip code, and gender) within a subpopulation to protect the identity of a patient. However, in order to enjoy a high-quality medical primary diagnosis service, the user's query data always contain personal physiological data, such as age, weights, and blood types, or even some ultimate personal identifiable information

such as fingerprints and DNA profiles. Once the nontrusted server in diagnosis system obtains the medical data, it may be able to identify an individual user easily [22], [23]. Hence, the anonymization mechanisms are not quite suitable for protecting the user's privacy in online medical primary diagnosis system. Differential privacy has become the de facto standard for privacy-preserving data analytics [24], [25], the central idea is to adequately obfuscate a query response by adding noise typically drawn from a Laplace distribution, such that the presence or absence of any user in the database is protected. However, these randomization approaches are often unsuitable for medical primary diagnosis, as they distort the data making it unusable for critical inferences, especially for physiological data, which is extremely strict about accuracy to avoid misdiagnosis. Different homomorphic encryption techniques are introduced in the medical diagnosis system [26]–[28], which enabled the healthcare SPs to process the encrypted query without gaining any knowledge on user's medical data, and the corresponding medical instruction without revealing any knowledge about the diagnosis system. However, the overhead of computation would be a stumbling block in making homomorphic encryption technology popularization in medical primary diagnosis system.

Different from all of the aforementioned works, our proposed CINEMA framework based on a skyline diagnosis model, which has a high accuracy. Moreover, aims at the efficiency and privacy issues, the CINEMA can protect users' medical data privacy and ensure the confidentiality of diagnosis model. Furthermore, based on fast secure permutation and comparison techniques, our proposed CINEMA can be easily implemented in smart terminals due to its high efficiency.

VIII. Conclusion

In this paper, we have introduced a precise diagnosis model by using skyline computation over the medical dataset, then, due to security and privacy concerns, a CINEMA framework, has been proposed. Within CINEMA, users can access online medical primary diagnosing service accurately without divulging their medical data. Detail analysis show that the CINEMA can keep the users' health information secret and the healthcare provider's diagnosis model confidential, with significantly less computation. In addition, performance evaluations via implementing CINEMA demonstrated its effectiveness and low communication overhead in term of the real environment.

Availability

The implementation of the proposed framework and relevant information can be downloaded at `https://www.xdzhuhui.com/demo/CINEMA/`.

References

[1] N. Tkachenko *et al.*, "Google trends can improve surveillance of type 2 diabetes," *Sci. Rep.*, vol. 7, p. 4993, Jul. 2017.

[2] M. Kay, J. Santos, and M. Takane, "mHealth: New horizons for health through mobile technologies," *World Health Organ.*, vol. 64, no. 7, pp. 66–71, 2011.

[3] *mHealth for Development: The Opportunity of Mobile Technology for Healthcare in the Developing World*, Vodafone Foundation, Washington, DC, USA, 2009.

[4] M. Sajid *et al.*, "All-printed highly sensitive 2D MoS_2 based multi-reagent immunosensor for smartphone based point-of-care diagnosis," *Sci. Rep.*, vol. 7, p. 5802, Jul. 2017.

[5] X. Liu, D. Yang, M. Ye, and W. Lee, "U-Skyline: A new skyline query for uncertain databases," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 945–960, Apr. 2013.

[6] S. Borzsony, D. Kossmann, and K. Stocker, "The skyline operator," in *Proc. IEEE 17th Int. Conf. Data Eng.*, Heidelberg, Germany, 2001, pp. 421–430.

[7] X. Han, J. Li, D. Yang, and J. Wang, "Efficient skyline computation on big data," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 11, pp. 2521–2535, Nov. 2013.

[8] D. Papadias, Y. Tao, G. Fu, and B. Seeger, "Progressive skyline computation in database systems," *ACM Trans. Database Syst.*, vol. 30, no. 1, pp. 41–82, 2005.

[9] T. Mingjie, Y. Yu, W. G. Aref, Q. Malluhi, and M. Ouzzani, "Efficient parallel skyline query processing for high-dimensional data," *IEEE Trans. Knowl. Data Eng.*, to be published.

[10] K. Hose and A. Vlachou, "A survey of skyline processing in highly distributed environments," *Int. J. Very Large Data Bases*, vol. 21, no. 3, pp. 359–384, 2012.

[11] X. Liu, R. Lu, J. Ma, L. Chen, and H. Bao, "Efficient and privacy-preserving skyline computation framework across domains," *Future Gener. Comput. Syst.*, vol. 62, pp. 161–174, Sep. 2016.

[12] Y. Park, J.-K. Min, and K. Shim, "Parallel computation of skyline and reverse skyline queries using mapreduce," *Proc. VLDB Endowment*, vol. 6, no. 14, pp. 2002–2013, 2013.

[13] W. Chen, M. Liu, R. Zhang, Y. Zhang, and S. Liu, "Secure outsourced skyline query processing via untrusted cloud service providers," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, San Francisco, CA, USA, 2016, pp. 1–9.

[14] J. Liu, J. Yang, L. Xiong, and J. Pei, "Secure skyline queries on cloud platform," in *Proc. IEEE 33rd Int. Conf. Data Eng. (ICDE)*, San Diego, CA, USA, 2017, pp. 633–644.

[15] N. Saleheen *et al.*, "mSieve: Differential behavioral privacy in time series of mobile sensor data," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Heidelberg, Germany, 2016, pp. 706–717.

[16] R. Whittaker, "Issues in mHealth: Findings from key informant interviews," *J. Med. Internet Res.*, vol. 14, no. 5, pp. 122–130, 2012.

[17] H. Lin, J. Shao, C. Zhang, and Y. Fang, "CAM: Cloud-assisted privacy preserving mobile health monitoring," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 6, pp. 985–997, Jun. 2013.

[18] H. Zhu, X. Liu, R. Lu, and H. Li, "Efficient and privacy-preserving online medical prediagnosis framework using nonlinear SVM," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 3, pp. 838–850, May 2017.

[19] X. Liu, H. Zhu, R. Lu, and H. Li, "Efficient privacy-preserving online medical primary diagnosis scheme on naive Bayesian classification," *Peer Peer Netw. Appl.*, vol. 11, no. 2, pp. 334–347, 2018.

[20] L. Sweeney, "k-anonymity: A model for protecting privacy," *Int. J. Uncertainty Fuzziness Knowl. Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[21] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Disc. Data*, vol. 1, no. 1, p. 52, 2007.

[22] M. Hansen, A. Schwartz, and A. Cooper, "Privacy and identity management," *IEEE Security Privacy*, vol. 6, no. 2, pp. 38–45, Mar./Apr. 2008.

[23] S. Al-Fedaghi and A. Al-Azmi, "Experimentation with personal identifiable information," *Intell. Inf. Manag.*, vol. 4, no. 4, pp. 123–133, 2012.

[24] C. Dwork, "Differential privacy: A survey of results," in *Proc. Int. Conf. Theory Appl. Models Comput.*, 2008, pp. 1–19.

[25] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference*. Heidelberg, Germany: Springer, 2006, pp. 265–284.

[26] Y. Rahulamathavan, S. Veluru, R. C.-W. Phan, J. A. Chambers, and M. Rajarajan, "Privacy-preserving clinical decision support system using Gaussian kernel-based classification," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 1, pp. 56–66, Jan. 2014.

[27] X. Liu, R. Lu, J. Ma, L. Chen, and B. Qin, "Privacy-preserving patient-centric clinical decision support system on naive Bayesian classification," *IEEE J. Biomed. Health Inform.*, vol. 20, no. 2, pp. 655–668, Mar. 2016.

[28] Y. Rahulamathavan, R. Phan, S. Veluru, K. Cumanan, and M. Rajarajan, "Privacy-preserving multi-class support vector machine for outsourcing the data classification in cloud," *IEEE Trans. Dependable Secure Comput.*, vol. 11, no. 5, pp. 467–479, Sep./Oct. 2014.

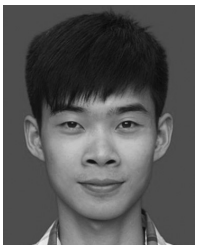
[29] D. Boneh and M. Franklin, "Identity-based encryption from the Weil pairing," in *Annual International Cryptology Conference*, Heidelberg, Germany: Springer, 2001, pp. 213–229.
[30] J. Yuan, Q. Ye, H. Wang, and J. Pieprzyk, *Secure Computation of the Vector Dominance Problem (International Conference on Information Security Practice and Experience)*. Heidelberg, Germany: Springer, 2008, pp. 319–333.
[31] D. Boneh, B. Lynn, and H. Shacham, "Short signatures from the Weil pairing," in *International Conference on the Theory and Application of Cryptology and Information Security*. Heidelberg, Germany: Springer, 2001, pp. 514–532.
[32] C. J. Merz and P. M. Murphy, *UCI Repository of Machine Learning Databases*, 1998.
[33] K. Mullesgaard, J. L. Pederseny, H. Lu, and Y. Zhou, "Efficient skyline computation in mapreduce," in *Proc. EDBT*, 2014, pp. 37–48.
[34] D. Kossmann, F. Ramsak, and S. Rost, "Shooting stars in the sky: An online algorithm for skyline queries," in *Proc. 28th Int. Conf. Very Large Databases (VLDB)*, Hong Kong, 2002, pp. 275–286.
[35] R. Lu, H. Zhu, X. Liu, J. K. Liu, and J. Shao, "Toward efficient and privacy-preserving computing in big data era," *IEEE Netw.*, vol. 28, no. 4, pp. 46–50, Jul./Aug. 2014.

Jiafeng Hua received the B.Sc. degree from the North University of China, Taiyuan, China, in 2012. He is currently pursuing the Ph.D. degree at the School of Cyber Engineering, Xidian University, Xi'an, China.

His current research interests include applied cryptography, cyber security, and privacy.

Hui Zhu (M'13) received the B.Sc. degree from Xidian University, Xi'an, China, in 2003, the M.Sc. degree from Wuhan University, Wuhan, China, in 2005, and the Ph.D. degree from Xidian University, in 2009.

He was a Research Fellow with the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore, in 2013. Since 2016, he has been a Professor with the School of Cyber Engineering, Xidian University. His current research interests include applied cryptography, data security, and privacy.



Fengwei Wang received the B.Sc. degree from Xidian University, Xi'an, China, in 2016, where he is currently pursuing the master's degree at the School of Cyber Engineering.

His current research interests include applied cryptography, cyber security, and privacy.

Ximeng Liu (S'13–M'16) received the B.Sc. degree in electronic engineering and Ph.D. degree in cryptography from Xidian University, Xi'an, China, in 2010 and 2015, respectively.

He is currently a Research Fellow with the School of Information System, Singapore Management University, Singapore, and a Qishan Scholar with the College of Mathematics and Computer Science, Fuzhou University, Fuzhou, China. His current research interests include cloud security, applied cryptography, and big data security.

Rongxing Lu (S'09–M'10–SM'15) received the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, in 2012.

He has been an Assistant Professor with the Faculty of Computer Science, University of New Brunswick, Fredericton, NB, Canada, since 2016. He was an Assistant Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, from 2013 to 2016. He was a Post-Doctoral Fellow with the University of Waterloo, from 2012 to 2013. His current research interests include applied cryptography, privacy enhancing technologies, and IoT-big data security and privacy.

Dr. Lu was a recipient of the Governor General's Gold Medal Award from the University of Waterloo and the 8th IEEE Communications Society (ComSoc) Asia–Pacific Outstanding Young Researcher Award in 2013. He currently serves as the Secretary of IEEE ComSoc CIS-TC.

Hao Li received the M.B., M.M., and M.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 2004, 2006, and 2014, respectively.

In 2016, she was with the Weil Institute of Emergency and Critical Care Research, Virginia Commonwealth University, Richmond, VA, USA, as a Research Fellow. Since 2017, she has been an Associate Professor with the First Affiliated Hospital, Xi'an Jiaotong University. Her current research interests include medical data analysis and critical care medicine.

Yeping Zhang is currently pursuing the bachelor's degree at the School of Cyber Engineering, Xidian University, Xi'an, China.

His current research interests include applied cryptography, cyber security, and privacy.