7-2020

CAMPS: Efficient and privacy-preserving medical primary diagnosis over outsourced cloud

Jianfeng HUA

Guozhen SHI

Hui ZHU

Fengwei WANG

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

See next page for additional authors

Published in Information Sciences, 2019 July, 527, 560-575.
https://doi.org/10.1016/j.ins.2018.12.054

CAMPS: Efficient and privacy-preserving medical primary diagnosis over outsourced cloud

Jiafeng Hua[a], Guozhen Shi[b], Hui Zhu[a,*], Fengwei Wang[a], Ximeng Liu[c], Hao Li[d]

[a] National Key Laboratory of Integrated Networks Services, Xidian University, China
[b] School of Information Security, Beijing Electronic Science and Technology Institute, China
[c] School of Information Systems, Singapore Management University, Singapore
[d] The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China

Keywords:
Medical primary diagnosis
Privacy-preserving
Skyline computation
Efficiency

A B S T R A C T

With the flourishing of ubiquitous healthcare and cloud computing technologies, medical primary diagnosis system, which forms a critical capability to link big data analysis technologies with medical knowledge, has shown great potential in improving the quality of healthcare services. However, it still faces many severe challenges on both users' medical privacy and intellectual property of healthcare service providers, which deters the wide adoption of medical primary diagnosis system. In this paper, we propose an efficient and privacy-preserving medical primary diagnosis framework (CAMPS). Within CAMPS framework, the precise diagnosis models are outsourced to the cloud server in an encrypted manner, and users can access accurate medical primary diagnosis service timely without divulging their medical data. Specifically, based on partially decryption and secure comparison techniques, a special fast secure two-party vector dominance scheme over ciphertext is proposed, with which CAMPS achieves privacy preservation of user's query and the diagnosis result, as well as the confidentiality of diagnosis models in the outsourced cloud server. Through extensive analysis, we show that CAMPS can ensure that users' medical data and healthcare service provider's diagnosis model are kept confidential, and has significantly reduce computation and communication overhead. In addition, performance evaluations via implementing CAMPS demonstrate its effectiveness in term of the real environment.

1. Introduction

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 and lowering the healthcare cost [29,35]. In medical primary diagnosis system, the user can deploy portable sensors around body to collect various physiological data, such as Electrocardiogram (ECG/EKG), blood pressure (BP), peripheral oxygen saturation (SpO2) and blood glucose [32]. These physiological data will be delivered to a central healthcare server for primary medical diagnosis via smart terminals, and the diagnosis result will be reported to the user and his doctor for decision making, which has great significance for healthcare monitoring and disease prevention at an early stage.

* Corresponding author.
E-mail address: zhuhui@xidian.edu.cn (H. Zhu).

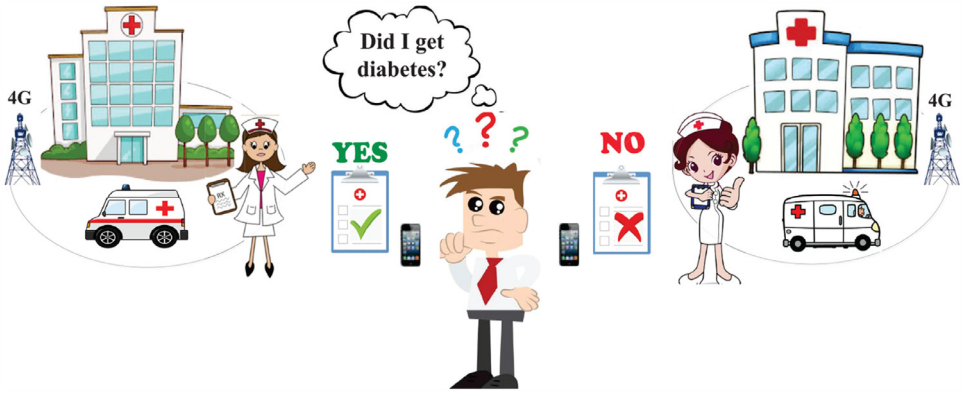


Fig. 1. Problem in medical primary diagnosis system.

Although medical primary diagnosis system provides medical instruction for patients anytime anywhere, while the problem "quality of diagnosis" may be the main stumble in blocking this technology in reality [10]. As shown in Fig. 1, a man delivers his physiological data to two medical servers (such as hospital) for primary medical diagnose via smart terminals, while the returned two completely different diagnosis result confused him. Since there exists a large body of prior works on medical primary diagnosis system [4,9,19], there are two main factors lead to medical diagnosis errors. First, the huge amounts of medical data collected by different healthcare server are too complex and voluminous to storage centralized for data analysing [9]. Moreover, the diagnosis models are generated by different healthcare server via different data mining technologies (such as Bayesian, neural network, or fuzzy logic theories), which are hardly to merge together [4]. Second, the diagnosis models are valuable asset for healthcare provider, neither party is willing to divulge any information to untrusted entities [19]. Therefore, these above factors place constraint on the mechanisms that can be used to generate diagnosis model in a distributed environment, while protecting the privacy of medical data.

As a powerful tool for multicriteria data analysis, data mining, and decision making, skyline query returns a set of interesting points which are the best trade-offs between the different dimensions of a huge data space [23]. 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 distributed database [13,18]. Specifically, data owners who store a fraction of available data prefer to performing cooperate data analysis to provide a more precise services by linking one or more databases, due to the additivity of skyline operator, skyline query can be executed in parallel to get the final skyline sets by merging the skyline candidates generated from the individual databases. Moreover, the physiological data has a standard reference region and extremely rich in information with high dimension [33], and the single aggregated distance metric with all dimensions is always hard to define, which may be quite appropriate for skyline query to applicated in medical data decision making [20].

To achieve low computational cost and convenient data process, healthcare providers often outsource their diagnosis model to a cloud server, which will handle users' medical queries by counting on its great computation power. 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 [36]. Generally, users are reluctant to send their health information directly to untrusted cloud server to obtain medical instruction. Meanwhile, the diagnosis model is also private and valuable asset, the healthcare providers are also unwillness to reveal any information about it to the cloud server [38]. Therefore, how to protect the privacy of users' medical data and the confidentiality of diagnosis model is crucially. Traditional anonymization techniques such as k-anonymity [34] and l-diversity [25] may be not quite suitable for protecting the user's privacy, due to the user's medical query 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 [1,27]. On the other hand, differential privacy has become the de facto standard for privacy-preserving data analytics [11,12], 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 [22,31], 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 multiple distributed medical datasets. Then, due to security and privacy concerns, we propose an efficient and privacy-preserving medical primary diagnosis framework (CAMPS). Within CAMPS, the precise diagnosis model is outsourced to the cloud server to provide medical primary diagnosis service in an encrypted manner, and users can access accurate medical primary diagnosis service timely without divulging their medical data. Specifically, the main contributions of this paper are as fourfold.

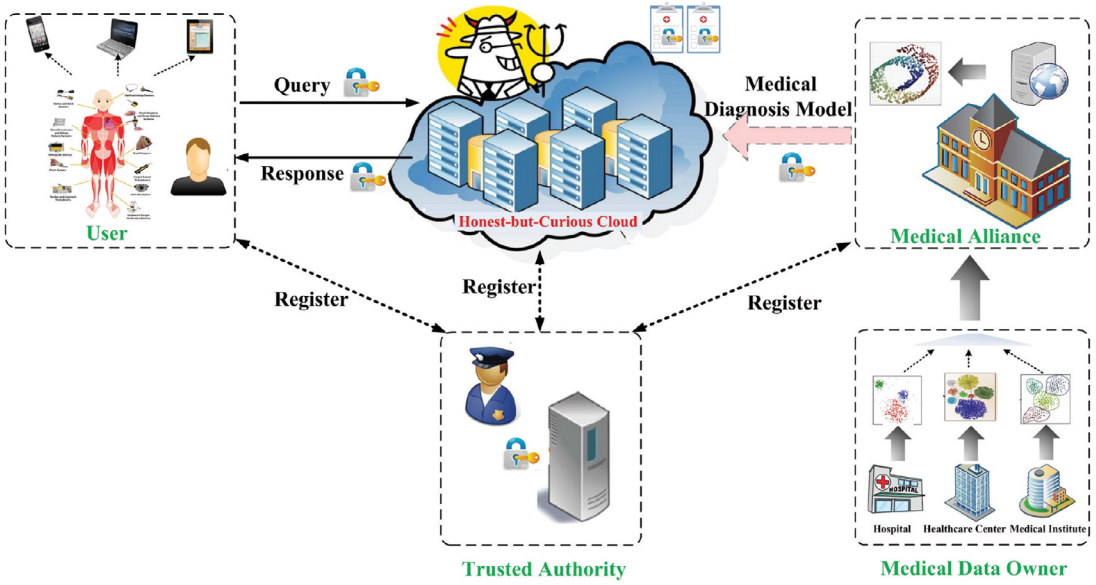


Fig. 2. System model under consideration.

- First, the proposed skyline diagnosis model is more precise. Different from traditional medical diagnosis model, which generates the diagnosis model from the individual medical database, the proposed skyline diagnosis model is generated by merging the skyline candidates output from several medical databases.
- Second, the proposed CAMPS is secure and privacy-preserving both for users' medical data and healthcare providers' diagnosis model. Within CAMPS, the encrypted user's query is directly operated at the cloud server without decryption, and the diagnosis result can only be decrypted by the user. Meanwhile, the diagnosis model in cloud provider can also be kept in an encrypted manner during the process.
- Third, the proposed CAMPS can provide the medical primary diagnosis service with high accuracy. Although the user's query request and 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 CAMPS can achieve high accuracy.
- Fourth, the proposed CAMPS is efficient in terms of computation and communication overhead. We have developed a custom simulator and implement CAMPS in a real environment, the performance evaluation demonstrates that our proposed CAMPS can provide efficient 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 2. In Section 3, we briefly review the skyline computation and additive homomorphic cryptosystem. Then, we introduce a novel diagnosis model by using skyline computation and propose our CAMPS framework in Section 4, followed by the security analysis and performance evaluation in Sections 5 and 6, respectively. We also review some related works in Section 7. Finally, we conclude this paper in Section 8.

2. System model, security requirement, and design goal

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

2.1. System model

In this work, we mainly focus on how to achieve precise and privacy-preserving diagnosis services over outsourced cloud server. Specifically, the system consists of five parts: *Trusted Authority* (TA), *Medical Dataowner* (MD), *User, Medical Alliance* (MA), and *Cloud Server* (CS), as shown in Fig. 2.

1) Trusted Authority: TA is an indispensable and trusted entity, who is in charge of distributing and managing all private keys in the medical primary diagnosis system.

2) Medical Dataowner: MD is considered to be a computation and storage limited entity (i.e., hospital, medical institute), who can generate a preliminary diagnosis model from individual private medical database. In order to provide high-quality medical diagnosis service, each MD delivers their preliminary diagnosis model to the MA for further cooperate analysis on their behalf.

3) Medical Alliance: MA is always considered as a trusted government organization or a management agent for MDs, and has a certain amount of computing and storage capabilities. In our system, MA is tasked to generate a final precise global diagnosis model based on the preliminary diagnosis models, which are collected from different MDs. Correspondingly, MA will shares certain level of mutual business interest with the participated MDs for incentives. With the advancement of cloud computing, MA tends to outsource the final global diagnosis model to the cloud server. Therefore, the MA mainly performs two functions: outsourcing diagnosis model to the cloud server and returning the diagnosis result back to users. With the process of outsourcing to the cloud server, the MA will perform some encryption operations to guarantee the diagnosis model's confidentiality, and performs partially decryption operations to obtain the final diagnosis result from the cloud server.

4) Cloud Server: CS has huge data storage space, and stores more than millions of encrypted precise diagnosis model from the MA and provides accurate medical query services for users. The cloud server also mainly performs two functions: authentication the users and computation over encrypted data. The authentication component is used to check users' identity, while the computing in encryption component is tasked to search and compute encrypted data items with users' encrypted query request. Furthermore, although the cloud server features high performance in computation and storage, since thousands of users may access query services at the same time, the efficiency of computation and communication are still challenging.

5) Users: Users who are registered in the MA can access the accurate medical diagnosis service from the precise medical diagnosis model that are outsourced to the cloud server. To guarantee the privacy of user's query which contains a lot of sensitive medical data collected by smart terminals, the user will perform some encryption operations before delivering it to the cloud server. Moreover, to lower energy costs, the encryption technique is required to be efficient and lightweight enough to adapt the resource constraint terminals.

2.2. Security requirements

The privacy of users' medical query and the confidentiality of MA's precise diagnosis model are crucial for the success of medical primary diagnosis system. In our security model, both the CS and users are considered to be honest-but-curious, while MA is trusted. Specifically, the MA generates the medical diagnosis model from distributed databases and keeps the diagnosis model secret from the MDs; CS strictly executes the protocol specifications to provide medical diagnosis service based on user's medical query, but it also tries to gain knowledge about the MA's diagnosis model and users' medical query for business benefit; users may intend to access 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: On one hand, the MA's medical diagnosis model are valuable assets should be kept secret from the cloud server, i.e., although the CS stores the medical diagnosis models and provides medical queries for users, it cannot gain any knowledge about medical data. On the other hand, the users' medical query should be protected from the CS, i.e., even if the CS obtains all queries from the user and corresponding responses from MA, it cannot identify the user's medical data accurately. Under this circumstance, the users' medical query and MA's medical primary diagnosis model can guarantee the privacy-preserving requirements. In addition, the privacy requirements also include the MA's responses can only be decrypted by legal users. It's worth note that, we do not consider any two parties from MA, CS, and users collude to disclose the third party's privacy in our current model. Moreover, the final precise medical diagnosis model generated by MA is kept secret from the MDs in our system model. Thus, the collusion attack on privacy is beyond the scope of this paper and will be discussed in our future research.

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

2.3. Design goal

Based on the aforementioned system model and security requirements, our design goal is to develop an efficient and privacy-preserving 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 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 medical primary diagnosis service, the designed privacy-preserving strategy cannot compromise the accuracy of diagnosis result. Therefore, the proposed framework should also achieve high accuracy.

3) Efficiency: Low communication overhead and computation complexity should be guaranteed. Considering the real time requirements of medical primary diagnosis service and the diversity of terminals, which might be constraint in resource (include computation, power, and storage, et al.), the proposed framework should achieve high communication and computation efficiency.

3. Preliminaries

In this section, we first review the definition of skyline computation, which serves as the basis of our proposed framework, then we introduce bilinear pairing technique [5], additive homormophic cryptosystem, and skyline computation[7].

3.1. 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 non-degenerated and efficiently computable bilinear map $\hat{e}$: $\mathbb{G} \times \mathbb{G} \longrightarrow \mathbb{G}_{\mathbb{T}}$ has the following properties.

1) Bilineariry: $\forall g, h \in \mathbb{G}$, 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 1. 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.

3.2. Additive homomorphic cryptosystem

Suppose that $[[m_1]]$ and $[[m_2]]$ are two additive homomorphic ciphertexts under the same public key pk in an additive homomorphic cryptosystem (e.g. Paillier cryptosystem [28]). The additive homomorphic cryptosystem has the additive homomorphism property:

$$[[m_1 + m_2]] = [[m_1]] + [[m_2]] \tag{1}$$

3.3. Skyline computation

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 2 (Positive Skyline Computation)**.** We define P_a positive dominated P_b, denoted by $Pdom(P_a, P_b)$, if it satisfies the following conditions: (1) $\forall 1 \le j \le m$, $P_a[j] \le P_b[j]$; (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 $Pdom(P_a, P_b)$ can be defined as

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

Definition 3 (Negative Skyline Computation)**.** We define P_a negative dominated P_b, denoted by $Ndom(P_a, P_b)$, if it satisfies the following conditions: (1) $\forall 1 \le j \le m$, $P_a[j] \ge P_b[j]$; (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 $Ndom(P_a, P_b)$ can be defined as

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

Definition 4 (Additivity of Skyline Computation)**.** According to reference [21], consider a medical dataset P and n datasets P_i such that $P = P_1 \cup \cdots \cup P_n$, the following equation holds:

$$SKY(P_1 \cup \cdots \cup P_n) = SKY(SKY(P_1) \cup \cdots \cup SKY(P_n)). \tag{4}$$

4. Proposed CAMPS framework

In this section, we first introduce a precise global skyline diagnosis model by merging several local skyline diagnosis models from distributed medical databases, then we propose our efficient and privacy-preserving medical primary diagnosis framework CAMPS, which consists of five phase: *1) system initialization; 2) data preparation; 3) query generation; 4) privacy-preserving medical primary diagnosis service; 5)query result reading*. Specifically, MA first provides registration for the user in the *system initialization* phase and executes some preprocess method on the global diagnosis model in the *data preparation* phase, and deliver the processed diagnosis model to the CS. Then, the user preprocesses the medical query by performing encryption operations in the *query generation* phase. After that, CS and MA perform the medical diagnosis service cooperatively with Paillier homomorphic technique in the *privacy-preserving online medical primary diagnosis service* phase. Finally, the user obtains the final diagnosis result from MA in the *query result reading* phase. The overall procedure of CAMPS was shown in Fig. 3. Meanwhile, we give the description of variables used in the following subsections in Table 1 for easier expression.

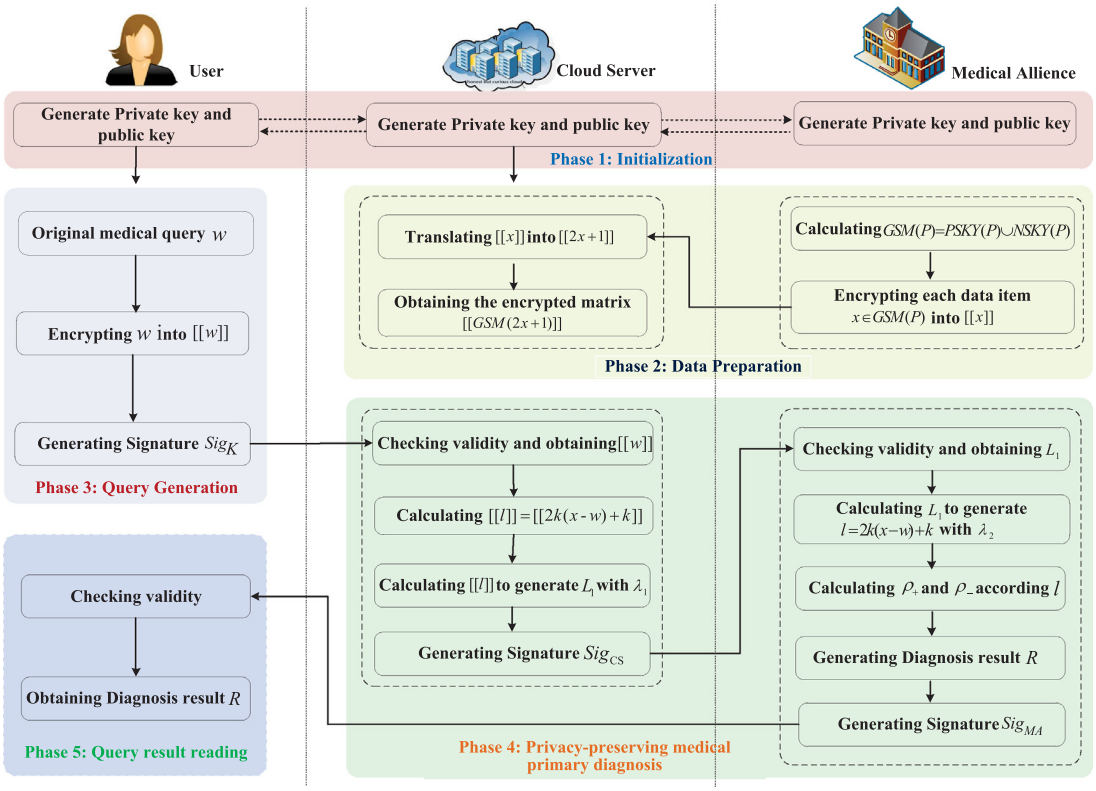


Fig. 3. The architecture of CAMPS.

Table 1
Definition of notations in CAMPS.

Notation	Definition
$g_1, q_1, q_2, n_1, h, \mathbb{G}, \mathbb{G}_T, \hat{e}$	The parameters of bilinear paring.
λ_1, λ_2	The partial of private key generated by TA.
$[[\cdot]]$	Encryption with public parameter (N, g).
$GSM(P)$	Global skyline diagnosis model generated from dataset P.
w, x, l	Vector with m-dimension.
$\hat{w}, \hat{x}$	The expanded vector with m-dimension.
L, U, V	Matrix with m-column and n-row.
ρ_+, ρ_-	Parameter of the diagnosis model.
$H(\cdot)$	Cryptographic hash function.
p, q, g, λ, N	The parameter of paillier cryptosystem.
r_1, r_2, r_3	Random number chosen by MA, CS, and U_i.
s, t	The number of vectors in $PSKY(P)$ and $NSKY(P)$.
$\|k\|$	The length of k.

4.1. Global skyline diagnosis model

Assuming there are several distributed medical dataset $P_1, \ldots, P_n$ in m-dimensional space, and a large medical dataset $P = P_1 \cup \cdots \cup P_n$. In order to generate the global skyline diagnosis model $GSM(P)$ from the medical dataset P, according to the definition 2, definition 3 and Eq. (4), we calculates the skyline candidates in $GSM(P)$ with the additivity of skyline computation:

$$\begin{aligned} GSM(P) &= PSKY(P) \cup NSKY(P) \\ &= PSKY(P_1 \cup \cdots \cup P_n) \cup NSKY(P_1 \cup \cdots \cup P_n) \\ &= PSKY(PSKY(P_1) \cup \cdots \cup PSKY(P_n)) \cup \\ &\quad NSKY(NSKY(P_1) \cup \cdots \cup NSKY(P_n)). \end{aligned}$$

From the above proof, we can conclude that due to the additivity of skyline computation, its easy to generate the global skyline diagnosis model in a distributed environment. Specifically, each sub-dataset can first operate skyline computation

on its dataset to extract the local skyline diagnosis model, then these local skyline diagnosis models will be delivered to the MA for merging the global skyline diagnosis model. Its worth noting that the local skyline diagnosis model is always not equal to the subset of the global skyline diagnosis model.

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 Eqs. (2) and (3), the skyline diagnosis standard is defined as follows:

$$\begin{cases} \rho_+ = \frac{1}{s}\sum_{i=1}^{s} Pdom(\vec{a_i}, \vec{w}) \\ \rho_- = \frac{1}{t}\sum_{j=1}^{t} Ndom(\vec{b_j}, \vec{w}). \end{cases} \tag{5}$$

Only if $0 < \rho_+ \leqslant 1$ and $-1 \leqslant \rho_- < 0$, we confirms that the diagnosis result is positive.

4.2. System initialization

TA will bootstrap the whole medical primary diagnosis system and manage all private keys for all participants. Specifically, TA first chooses a security parameter k_1 and operates function $Gen(k_1)$ to generate the bilinear parameters $(p_1, q_1, g_1, \mathbb{G}, \mathbb{G}_\mathbb{T}, \hat{e}, h, n_1)$, where $n = p_1 \cdot q_1$. Then, TA chooses a random number in $\mathbb{Z}_q^*$ as its private key SK_{TA}, and computes its public key $PK_{TA} = {g_1}^{SK_{TA}}$. In addition, TA chooses a secure asymmetric encryption algorithm $E()$, i.e., ECC, and a secure cryptographic hash function $H()$, where $H : \{0, 1\}^* \longrightarrow \mathbb{G}$. Finally, TA keeps the tuple $\langle q_1, SK_{TA}\rangle$ as master key secretly, and publishes the system parameters $< n_1, g_1, h, \mathbb{G}, \mathbb{G}_\mathbb{T}, \hat{e}, PK_{TA}, E(), H() >$.

MA chooses a random number in $\mathbb{Z}_{q_1}^*$ as its private key SK_{MA}, and computing the corresponding public key $PK_{MA} = {g_1}^{SK_{MA}}$. Similarly, the cloud server CS and user U_i also choose random numbers in $\mathbb{Z}_{q_1}^*$ as their private key SK_{CS} and SK_{U_i}, then computing their public key $PK_{CS} = {g_1}^{SK_{CS}}$ and $PK_{U_i} = {g_1}^{SK_{U_i}}$. TA generates two large prime numbers p, q ($N = p \cdot q$, $|p| = |q|$) and a generator g of order $(p-1)(q-1)/2$ as the public parameter, then spit the private key λ into two parts λ_1 and λ_2, where $\lambda = lcm(p-1, q-1)$. When the cloud server, MA, and U_i registering in the TA, TA sends (N, λ_1) back to the cloud server, (N, λ_2) to the MA, and N to the U_i through a secure channel.

4.3. Data preparation

The global diagnosis model $GSM(P)$ has plenty of medical data points, which contains lots of medical attributes (such as age, blood pressure, heart rate, et al.). In general, these medical data points are stored in MA with plaintext format, which can be presented as lots of vectors. Due to $GSM(P) = PSKY(P) \cup NSKY(P)$, the global diagnosis model $GSM(P)$ can be shown as follow:

$$GSM(P) = \begin{pmatrix} \vec{a_1} \\ \cdots \\ \vec{a_s} \\ \vec{b_1} \\ \cdots \\ \vec{b_t} \end{pmatrix} = \begin{vmatrix} a_{11} & , \cdots, & a_{1m} \\ \cdots & , \cdots, & \cdots \\ a_{s1} & , \cdots, & a_{sm} \\ b_{11} & , \cdots, & b_{1m} \\ \cdots & , \cdots, & \cdots \\ b_{t1} & , \cdots, & b_{tm} \end{vmatrix}.$$

Before being uploaded to the cloud server, all vectors in global skyline diagnosis model should be blurred for security reasons, the procedure as follows.

- For each data item $x_{ij} \in GSM(P)$, where $1 \le i \le n$ and $1 \le j \le m$, MA first chooses a random number $r_1 \in \mathbb{Z}_{N^2}^*$, then performs the encryption operation as

$$[\![x_{ij}]\!] = (1 + x_{ij}N) \cdot g^{r_1} \ mod \ N^2. \tag{6}$$

- The MA computes the blurred global skyline diagnosis model $[[GSM(P)]]$ with a secure asymmetric encryption algorithm $E(\cdot)$ and the public key PK_{MA}, then outsourced the encrypted $[[GSM(P)]]$ to the cloud server.

After receiving the encrypted global skyline diagnosis model $GSM(P)$ from MA, the cloud server mainly performs decryption and homomorphic operations.

- The cloud server first obtains the blurred global skyline diagnosis model $[[GSM(P)]]$ with the secret key SK_{CS}.
- For each data item $[[x_{ij}]] \in [[GSM(P)]]$, the cloud server chooses a random number $r_2 \in \mathbb{Z}_{N^2}^*$ and performs the following computations with Eqs. (1) and (6).

$$[\![\hat{x}_{ij}]\!] = [\![x_{ij}]\!]^2 \cdot [\![1]\!] = [\![x_{ij}]\!] \cdot [\![x_{ij}]\!] \cdot [\![1]\!]$$

$$
\begin{aligned}
&= (1 + 2x_{ij} \cdot N) \cdot g^{2r_1} \cdot (1 + N) \cdot g^{r_2} \ mod\ N^2 \\
&= (1 + (2x_{ij} + 1) \cdot N) \cdot g^{(2r_1 + r_2)} \ mod\ N^2 \\
&= [\![2x_{ij} + 1]\!]. \qquad (7)
\end{aligned}
$$

All the data items in $[[GSM(P)]]$ can be presented as follows:

$$
[\![GSM(\hat{P})]\!] = \begin{vmatrix}
[\![2a_{11} + 1]\!] & ,\cdots, & [\![2a_{1m} + 1]\!] \\
\cdots & ,\cdots, & \cdots \\
[\![2a_{s1} + 1]\!] & ,\cdots, & [\![2a_{sm} + 1]\!] \\
[\![2b_{11} + 1]\!] & ,\cdots, & [\![2b_{1m} + 1]\!] \\
\cdots & ,\cdots, & \cdots \\
[\![2b_{t1} + 1]\!] & ,\cdots, & [\![2b_{tm} + 1]\!]
\end{vmatrix}. \qquad (8)
$$

4.4. Query generation

After registering in the MA, user U_K wants to access medical primary diagnosis service from the cloud server, before sending the query request, there are some blurred operations should be performed for privacy concerns.

- Assume the query request can be presented as a vector $\vec{w} = (w_1, \ldots, w_m) \in \mathbb{Z}_N$, for each data item $w_j \in w$, where $1 \le j \le m$, U_K chooses a random number $r_3 \in \mathbb{Z}^*_{N^2}$ and performs the encryption operation as Eq. (6).

$$
[\![w_j]\!] = (1 + w_j N) \cdot g^{r_3} \ mod\ N^2. \qquad (9)
$$

- Let $Q_{U_K} = < U_K || [\![w]\!] || TS_1 >$, where $[\![w]\!] = ([\![w_1]\!], \cdots, [\![w_m]\!])$ and TS_1 is the current timestamp, which is used to resist the potential replay attack. U_K generates a signature $Sig_K = (H(Q_{U_K}))^{SK_{U_K}}$ with his/her private key SK_{U_K}, and computes the medical query request $E_{Q_{U_K}} = E_{PK_{CS}}(Q_{U_K} || Sig_K)$ with the cloud server's public key PK_{CS}, then send it to the cloud server.

4.5. Privacy-preserving diagnosis and response

After receiving $E_{Q_{U_K}}$, the cloud server verifies its validity firstly, then performs some computation and partially decryption operations on the medical query.

- The cloud server decrypts $E_{Q_{U_K}}$ with its secret key SK_{CS} to obtain Q_{U_K} and Sig_K, then verifies its validity by checking whether $\hat{e}(g_1, Sig_K) = \hat{e}(PK_{U_K}, H(Q_{U_K}))$. If it does hold, the received medical query request $E_{Q_{U_K}}$ is valid. Then the cloud server extracts the blurred medical query vector $[[w]]$.

- For each data item in $[[w_j]] \in [[w]]$, the cloud server performs the following operations with Eq. (9).

$$
\begin{aligned}
[\![\hat{w}_j]\!] &= [\![w_j]\!]^2 = [\![w_j]\!] \cdot [\![w_j]\!] \\
&= (1 + 2w_j \cdot N) \cdot g^{2r_3} \ mod\ N^2 \\
&= [\![2w_j]\!]. \qquad (10)
\end{aligned}
$$

- For each data item $[\![\hat{x}_{ij}]\!]$, where $1 \le i \le n$ and $1 \le j \le m$, according Eqs. (7) and (10), the cloud server choose a random number $k \in \mathbb{Z}_N$ to compute

$$
\begin{aligned}
[\![l_{ij}]\!] &= ([\![\hat{x}_{ij}]\!] \cdot [\![\hat{w}_j]\!]^{N-1})^k \\
&= ([\![\hat{x}_{ij}]\!] \cdot ((1 + (N-1) \cdot 2w_j \cdot N) \cdot g^{2r_3 \cdot (N-1)} \ mod\ N^2))^k \\
&= ([\![\hat{x}_{ij}]\!] \cdot ((1 - 2w_j \cdot N) \cdot g^{2r_3 \cdot (N-1)} \ mod\ N^2))^k \\
&= ([\![2x_{ij} + 1]\!] \cdot [\![-2w_j]\!])^k \\
&= ((1 + (2x_{ij} - 2w_j + 1) \cdot N) \cdot g^{2r_1 + r_2 + 2(N-1)r_3})^k \\
&= (1 + k(2x_{ij} - 2w_j + 1) \cdot N) \cdot g^{k(2r_1 + r_2 + 2(N-1)r_3)} \\
&= [\![2k(x_{ij} - w_j) + k]\!]. \qquad (11)
\end{aligned}
$$

According to the Eqs. (8) and (11), the matrix $[[L]]$ can be presented in the form as follows:

$$
[\![L]\!] = \begin{vmatrix}
[\![2k(a_{11} - w_1) + k]\!] & ,\cdots, & [\![2k(a_{1m} - w_m) + k]\!] \\
\cdots & ,\cdots, & \cdots \\
[\![2k(a_{s1} - w_1) + k]\!] & ,\cdots, & [\![2k(a_{sm} - w_m) + k]\!] \\
[\![2k(b_{11} - w_1) + k]\!] & ,\cdots, & [\![2k(b_{1m} - w_m) + k]\!] \\
\cdots & ,\cdots, & \cdots \\
[\![2k(b_{t1} - w_1) + k]\!] & ,\cdots, & [\![2k(b_{tm} - w_m) + k]\!]
\end{vmatrix}.
$$

- The cloud server performs partial decryption on matrix $[[L]]$ to compute U by using the part of private key $SK_{(1)} = \lambda_1$, for each $u_{ij} \in U$, where $1 \le i \le n$ and $1 \le j \le m$, we denote random number $r = k(2Nr_3 + 2r_1 + r_2 - 2r_3)$ and the computation process shown as follows:

$$\begin{aligned} u_{ij} &= [\![l_{ij}]\!]^{\lambda_1} \\ &= (1 + l_{ij}N)^{\lambda_1} \cdot g^{\lambda_1 r} \ mod \ N^2 \\ &= (1 + l_{ij}\lambda_1 \cdot N) \cdot g^{\lambda_1 r} \ mod \ N^2. \end{aligned} \tag{12}$$

- Let $Q_{CS} =< U||[\![L]\!]||TS_2 >$, where TS_2 is the current timestamp, which is used to resist the potential replay attack. The cloud server generates a signature $Sig_{CS} = (H(Q_{CS}))^{SK_{CS}}$ with his/her private key SK_{CS}, and computes $E_{Q_{CS}} = E_{PK_{MA}}(Q_{CS}||Sig_{CS})$ with MA's public key PK_{MA}, then send it to the MA.

Upon receiving $E_{Q_{CS}}$, the MA verifies its validity firstly, then performs partially decryption operations on the intermediate calculate result.

- The MA decrypts $E_{Q_{CS}}$ with its secret key SK_{MA} to obtain Q_{CS} and Sig_{CS}, then verifies its validity by checking whether $\hat{e}(g, Sig_{CS}) = \hat{e}(PK_{CS}, H(Q_{CS}))$. If it dose hold, the received intermediate calculate result $E_{Q_{CS}}$ is valid. Then the MA extracts the matrix U and $[[L]]$.

- The MA performs partial decryption on matrix $[[L]]$ to compute V by using the part of private key $SK_{(2)} = \lambda_2$ and Eq. (12), for each $v_{ij} \in V$, where $1 \le i \le n$ and $1 \le j \le m$, the computation shown as follows:

$$\begin{aligned} v_{ij} &= [\![l_{ij}]\!]^{\lambda_2} \cdot u_{ij} \\ &= (1 + l_{ij}\lambda_2 \cdot N) \cdot (1 + l_{ij}\lambda_1 \cdot N) \cdot g^{r(\lambda_1+\lambda_2)} \ mod \ N^2 \\ &= (1 + l_{ij}(\lambda_1 + \lambda_2) \cdot N) \cdot g^{r(\lambda_1+\lambda_2)} \ mod \ N^2 \\ &= (1 + l_{ij}\lambda \cdot N) \cdot g^{r\lambda} \ mod \ N^2 \\ &= (1 + l_{ij}\lambda \cdot N). \end{aligned} \tag{13}$$

Then, due to $gcd(\lambda, N) = 1$ and Eq. (13), MA recovers l_{ij} with the computation:

$$\begin{aligned} l_{ij} &= \left(\frac{(1 + l_{ij} \cdot N\lambda) - 1}{N} \ mod \ N^2 \right) \cdot \lambda^{-1} \ mod \ N \\ &= 2k(x_{ij} - w_j) + k. \end{aligned} \tag{14}$$

- Due to $x_{ij}, w_j \in \mathbb{Z}_q$, we confirms the relationship between $x_{ij} \in L$ and $w_j \in w$ from the value of l_{ij}. If $0 < l_{ij} < N/2$, then $x_{ij} \ge w_j$; otherwise, $x_{ij} < w_j$. According to Eqs. (2) and (3), MA obtains the dominance relationship of each dimension in $GSM(P)$ and the medical query vector w, and makes determination by performing the computations as follows.

$$\begin{cases} \rho_+ = \frac{1}{s} \sum_{i=1}^{s} Pdom(a_i, w) \\ \rho_- = \frac{1}{t} \sum_{j=1}^{t} Ndom(b_j, w). \end{cases}$$

Only if $0 < \rho_+ \leqslant 1$ and $-1 \leqslant \rho_- < 0$, we denote the diagnosis result $R = 1$, which means the diagnosis result is positive.

- Let $R_{MA} =< R||TS_3 >$, where TS_3 is the current timestamp, which is used to resist the potential replay attack. The MA generates a signature $Sig_{MA} = (H(R_{MA}))^{SK_{MA}}$ with his/her private key SK_{MA}, and computes $E_{R_{MA}} = E_{PK_{U_K}}(R_{MA}||Sig_{MA})$ with user's public key PK_{U_K}, then send it to the U_K.

4.6. Query result reading

Upon receiving $E_{R_{MA}}$, U_K verifies its validity firstly, then performs decryption operations to get the diagnosis result.

- U_K decrypts $E_{R_{MA}}$ with its secret key SK_{U_K} to obtain R_{MA} and Sig_{MA}, then verifies its validity by checking whether $\hat{e}(g, Sig_{MA}) = \hat{e}(PK_{MA}, H(R_{MA}))$. If it dose hold, the received response $E_{R_{MA}}$ is valid.

- Then U_K extracts the final diagnosis result R. If $R = 1$, it means that the diagnosis result is positive and he/she has got the certain disease. Otherwise, he/she is healthy.

Correctness. In Eq. (13), in order to ensure the correctness of the decryption, considering the aforementioned constraints, ie,. $N = pq$, where p and q are two large prime numbers and $|p| = |q|$, we set the generator $g = -a^{2N}$, where a is a random number satisfy $a \in \mathbb{Z}^*_{N^2}$. In Eq. (14), the length of k and N should satisfy the constraints: $|k| < |N|/4$, ie,. when $|N| = 512$, we just set $|k| = 100$.

5. Security analysis

In this section, we analyze the security properties of the proposed CAMPS framework. Specifically, following the security requirements discussed in Section 2, our security analysis will mainly focus on three parts: how the CAMPS framework protects the privacy of the user's medical query, ensures the confidentiality of global skyline diagnosis model, and authenticates the query request and response.

1) *The Privacy of User's Medical Query*. The user's medical query is privacy preserving during the full procedure in the proposed CAMPS framework.

- In *Query Generation* phase, the medical query w is encrypted by performing operations $[\![w_j]\!] = (1 + w_j N) \cdot g_3^r \ mod \ N^2$ for each $w_j \in w$ before being sent to the cloud server. After receiving the encrypted medical query $[[w]]$ in *Privacy-Preserving Diagnosis and Response* phase, the cloud server performs some computation operations on $[[w]]$ to generate an intermediate result $[\![2k(x_{ij} - w_j) + k]\!]$ for each $x_{ij} \in GSM(P)$, while the medical query $[[w]]$ is always in an encrypted format, and the generator g and N are only known by the MA and the registered users, therefore, the cloud server cannot obtain the user's real medical query w from $[[w]]$.

- In *Privacy-Preserving Diagnosis and Response* phase, the cloud server delivers the intermediate result $[\![2k(x_{ij} - w_j) + k]\!]$ to the MA for computing the final diagnosis result R. Although MA can obtain the real value of $2k(x_{ij} - w_j) + k$ by using the partially private key λ_2, the generator g, and N, since k is kept secret by the cloud server, MA can only obtain the dominance relationship between vectors x_i and w, while can not get the accurate information about the medical query w.

Due to the collusion attack between the cloud server and the MA is not considered, moreover, the communication between U_K, the cloud server, and MA is transmitted under secure channel, and only the valid entity can obtain the encrypted query request. Thus, the user's medical query is privacy-preserving during the full procedure in the proposed CAMPS framework.

2) *The Confidentiality of Diagnosis Model*. The proposed CAMPS framework can achieve confidential on MA's global skyline diagnosis model during the full procedure.

- In *Data Preparation* phase, since the global skyline diagnosis model consist by a lot of skyline data points, each data item is encrypted by performing operations $[\![x_{ij}]\!] = (1 + x_{ij} N) \cdot g^{r_1} \ mod \ N^2$ before being sent to the cloud server as well. After receiving the encrypted global skyline diagnosis model $[[GSM(P)]]$, the cloud server compute $[\![\hat{x}_{ij}]\!] = [\![2x_{ij} + 1]\!]$ at first. When the user U_K access the medical primary diagnosis service by sending the encrypted medical query $[[w]]$, the cloud server performs some computation operations on the each data item $[[x_{ij}]] \in [[GSM(P)]]$ to generate the intermediate result $[\![2k(x_{ij} - w_j) + k]\!]$, while the data item $[[x_{ij}]]$ is always in an encrypted format, and the generator g and N are only known by the MA and the registered users, therefore, the cloud server cannot obtain the MA's data item x_{ij} from $[[x_{ij}]]$.

- In *Privacy-Preserving Diagnosis and Response* phase, After receiving the intermediate result $[\![2k(x_{ij} - w_j) + k]\!]$, MA obtains the dominance relationship between each dimension $x_{ij} \in GSM(P)$ and w_j from the value of $2k(x_{ij} - w_j) + k$ by performing the final partially decryption with λ_2 at first, then generates the final diagnosis result R by performing the standard of skyline diagnosis model. When user U_K obtain the diagnosis result R in *Query Result Reading* phase, since the value of R is either 1 or 0, which means positive and negative in certain disease, therefore, U_K can not gain any accurate information about the data item $x_{ij} \in GSM(P)$.

Due to the collusion attack between the cloud server and the user is not considered, moreover, the response from MA to U_K is transmitted under secure channel, and only the valid user can obtain the encrypted response. Thus, the proposed CAMPS framework can achieve confidential on MA's global skyline diagnosis model during the full procedure.

3)*The Authentication of Query Request and Response*. In the proposed CAMPS framework, each registered user's request is signed by Boneh–Lynn–Shacham (BLS) short signature [6]. 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 the cloud server without the valid secret key, she/he also cannot submit valid query request to the cloud server. As a result, the query request from the unregistered user and response from the mendacious cloud server can be detected in the proposed CAMPS framework.

From the above analysis, we can conclude that the proposed CAMPS framework is secure and privacy-preserving both for user, the cloud server and MA, and all the security requirements are achieved as well.

6. Performance evaluations

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

(a) Select Organ. (b) Choose disease. (c) Medical data. (d) Diagnosis result.

Fig. 4. Implementation of CAMPS.

Table 2
Comparison of accuracy.

Accuracy	CAMPS	SRPD
Yes(538)	518(96%)	518(96%)
No(379)	365(96%)	365(96%)
Overall(917)	885(96%)	885(96%)

6.1. Evaluation environment

In order to measure the comprehensive performance of CAMPS in the real environment, we implement CAMPS on computers and smart phone with a real medical database. Specifically, a smart phone with 2.4 GHz eight-core processor, 4-GB RAM, and an HUAWEI EMUI 5.0 operating system is chosen to evaluate the medical user; Three computers with 2.3 GHz six-core processor, 16GB RAM, Windows 7, are chosen to evaluate the cloud server, MA and TA, respectively, which are connected through 802.11g WLAN. Based on CAMPS framework, we construct three simulators in the computer to simulate the cloud server, MA, and TA, another simulator in the smart phone to simulate the user. As shown in Fig. 4, a user choose the type of disease and input the value of each physiological data item through the client, then get the primary medical diagnosis result from the cloud server. In order to obtain the correct primary diagnosis result, we can just set the length of parameter $|r_1| = |r_2| = |r_3| = 512$. In addition, we consider one real dataset which is from the UCI machine learning repository called Heart Disease Data (HDD) Set [2] to evaluate the accuracy of our proposed framework.

6.2. Accuracy evaluation

Based on the HDD, we choose four different datasets such as cleveland.data (298 instances), hungarin.data (293 instances), long-beach-va.data (202 instances), and switzerland.data (124 instances) to evaluate the accuracy of CAMPS. Each item in the dataset 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 blood pressure 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 generating the skyline diagnosis model, all the instances from the HDD should be normalized. Then, each dataset generates its local skyline diagnosis model and computes the final global skyline diagnosis model cooperative with skyline computation. After that, we test the success rate in the plain domain (abbreviated as SRPD) by using the global skyline diagnosis model $GSM(P)$ and HDD. Meanwhile, we take advantage of $GSM(P)$ to evaluate the accuracy of our proposed CAMPS framework with same evaluation environment. Thus, we obtained the comparison of accuracy. As shown in Table 2, we can see that the total number of correctly diagnosed heart disease instances is 518 out of 538 and that of non-heart disease instances is 365 out of 379. In total, 885 samples are correctly classified out of 917(96%), our privacy-preserving framework does not compromise the accuracy, and the test result also confirms it by achieving the same accuracy as that of SRPD.

6.3. Computation complexity

For the proposed CAMPS framework, there are three parties (include users, the cloud server, and MA) involved in the computation to provide medical primary diagnosis service. Suppose the global skyline diagnosis model $GSM(P)$ contains n elements, each element has m attributes. We assume that one regular exponentiation operation with an exponent of length $|M|$ requires $1.5|M|$ multiplications [15] (e.g., if the length or r is $|M|$, then the computation of g^r is $1.5|M|$ multiplications).

Table 3
Comparison of computation complexity .

	User	Cloud server C_1	MA/Cloud server C_2
CAMPS	$1.5m\|M\|$	$(3mn+1.5m)\|M\|$	$1.5mn\|M\|$
FSSP [20]	$1.5m\|M\|$	$(3nl+(3n+1.5)\log_2 n+1.5nm+1.5l)\|M\|$	$(3nl+(1.5n+1.5)\log_2 n+1.5nm+1.5l)\|M\|$

As exponentiation operation is significantly more computation costly than the addition and multiplication operations, we ignore the fixed numbers of addition and multiplication operation in our analysis.

In the phase of *Query Generation*, the user performs operations $[\![w_j]\!] = (1+w_jN)\cdot g^{r_3} \ mod\ N^2$ for each $w_j \in w$ before being send to the cloud server, where $1 \le j \le m$, which requires $1.5m|M|$ multiplications. After receiving the encrypted medical query $[[w]]$ in *Privacy-Preserving Diagnosis and Response* phase, the cloud server first performs computation operations on $[[w_j]]$ to generate $[\![\hat{w}_j]\!]$, which requires $1.5m|M|$ multiplications; After that, for each element $[[x_{ij}]] \in [[GSM(P)]]$, where $1 \le i \le n$, the cloud server computes $([\![\hat{x}_{ij}]\!]\cdot[\![\hat{w}_j]\!]^{N-1})^k$ to generate $[\![l_{ij}]\!] = [\![2k(x_{ij}-w_j)+k]\!]$, which requires $1.5mn|M|$ multiplications; Last, the cloud server performs partial decryption on $[\![l_{ij}]\!]^{\lambda_1}$ to generate u_{ij} with the part of private key $SK(1)=\lambda_1$, which requires $1.5mn|M|$ multiplications. Thus, the cloud server cost approximate $(3mn+1.5m)|M|$ multiplications totally during the medical diagnosis service. When MA receives the U, it performs final partial decryption on $[[l_{ij}]]$ to compute $v_{ij} = [\![l_{ij}]\!]^{\lambda_2}\cdot u_{ij}$ by using the partially private key $SK(2)=\lambda_2$, then recovers $l_{ij} = 2k(x_{ij}-w_j)+k$ with the λ^{-1}, which requires $1.5mn|M|$ multiplications.

Different from other time-consuming encryption techniques, the proposed CAMPS framework achieves high accuracy medical primary diagnosis service and largely reduce the encryption time for the smartphone and the cloud server by using the paillier cryptosystem with threshold decryption technique. In order to compare with CAMPS, we select a privacy-preserving scheme which performs secure skyline queries over encrypted cloud server, we denoted it as FSSP [20]. Within FSSP, cloud server C_1 was tasked to storage the encrypted medical records and performs the main computation, an other non-colluding cloud server C_2 hold the private key shared by the data owner and assist with the computation, while the user dose not need to participate in any computation except encrypt the medical query and combine the partial result from the cloud server C_1 and C_2. We assume the dimension of query vector is m and the number of vectors in the cloud server C_1 is n, the corresponding computational costs of the user, the cloud server C_1 and C_2 are $1.5m|M|$ multiplications, $3nl+(3n+1.5)\log_2 n+1.5nm+1.5l$ multiplications and $3nl+(1.5n+1.5)\log_2 n+1.5nm+1.5l$ multiplications, where l is the length of the attributes in the vectors.

As shown in Table 3, it is obvious that our proposed CAMPS framework can achieve efficient medical primary diagnosis with low computation complexity both in the cloud server and MA, while the computation complexity of users is equal, because the client have not participated in the computation during the diagnosis. To further demonstrate the advantage of CAMPS, we denote the cloud server and MA as the healthcare service provider (SP) and evaluate the total average running time under the evaluation environment described in Section 6.1. Fig. 5 depict the computation overhead varying with the dimension of the query vector and the number of vectors in SP. Through comparing Fig. 5, we can find that with the increase of the numbers of vectors, the computation overhead of FSSP significantly increases and it is much higher than that of our proposed CAMPS framework. Although the computation overhead of our proposed CAMPS framework also increases when the number of vectors is large, it is still much lower than that of FSSP. In conclusion, our proposed CAMPS framework can achieve better efficiency on computation overhead in SP.

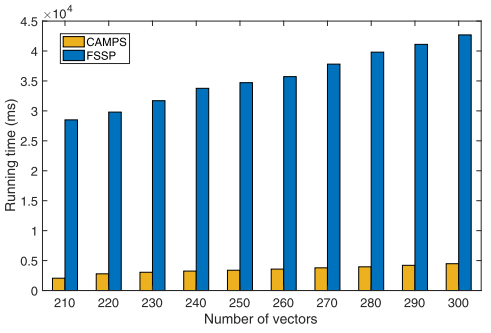
6.4. Efficiency evaluation

In order to test the factors that may affect the efficiency of our proposed CAMPS, different $GSM(P)$ are randomly generated. We evaluate the computation cost and communication overhead both in the cloud server, MA and user. Based on the definition of skyline computation, we can note that the dimensions of vectors and the total number of vectors in $GSM(P)$ may be the main factors that impact the computation complexity on the cloud server in CAMPS. Therefore, we choose different dimensions and number of vectors to illustrate the computation cost. The dimension is selected from 2 to 11, and the number is from 210 to 300. In order to ensure the accuracy, we perform the experiment 1000 times with different dimensions and numbers.

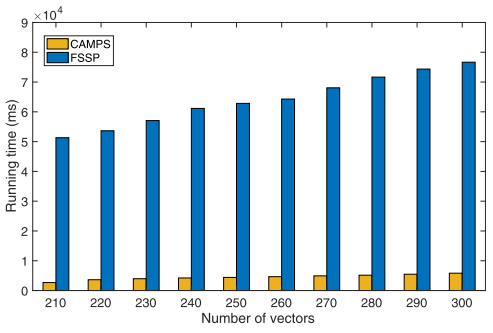
1) Computation Cost. We first evaluate the main factors that impact the computation cost of the cloud servers, the MA and user.

The Cloud Server: As shown in Fig. 6(a), we can learn that the computation overhead of the cloud server is increased with the dimension and number. When providing the medical primary diagnosis service, the cloud server have to compute the intermediate result $[\![2k(x_{ij}-w_j)+k]\!]$ and perform partially decryption operations on each $x_{ij} \in GSM(P)$, where $1 \le i \le n$ and $1 \le j \le m$, which cost much more time with the increase of vectors' dimension and number. However, due to the fact that basic operations are based on paillier cryptosystem with threshold decryption techniques, the maximum time required for the cloud server is less than 16s under the evaluation environment.

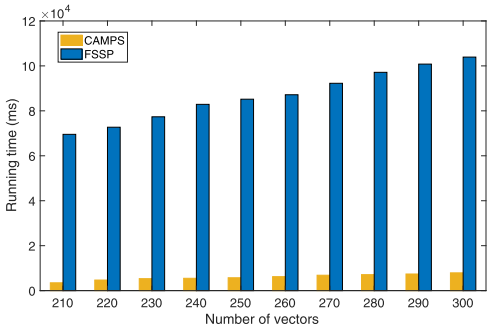
The MA: As shown in Fig. 6(b), we can also learn that the computation overhead of the MA is increased with the dimension and number. The reason is that, after receiving the intermediate result u_{ij} and $[l_{ij}]$ from the cloud server, MA have to



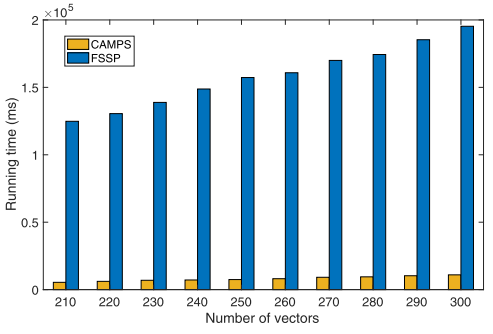
(a) The impact of n (m=4).



(b) The impact of n (m=6).

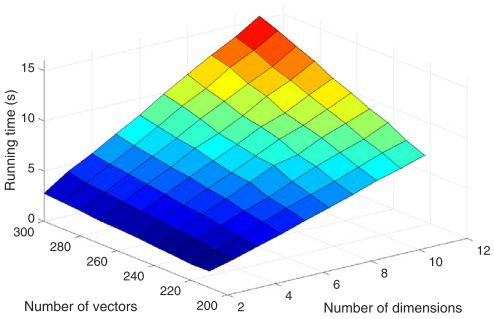


(c) The impact of n (m=8).

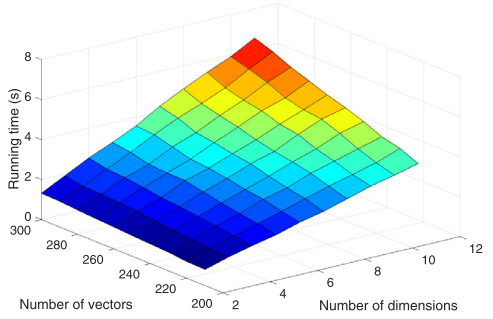


(d) The impact of n (m=10).

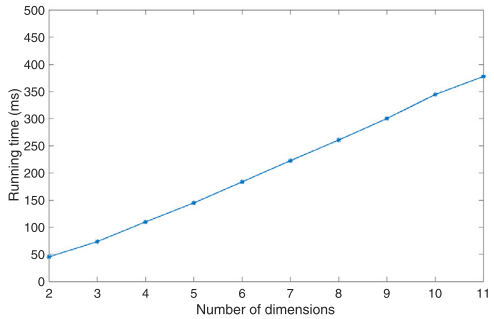
Fig. 5. Average running time in CAMPS vs FSSP ($K = 512$).



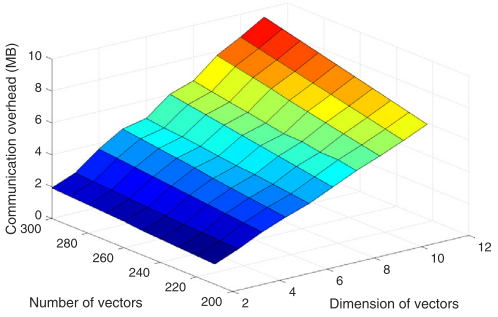
(a) Computation cost of cloud server.



(b) Computation cost of MA.



(c) Computation cost of user.



(d) Communication Cost.

Fig. 6. Performance evaluation of CAMPS.

perform the final partially decryption and multiplication operations on each u_{ij} and $[l_{ij}]$, the computation time also increased with vectors' dimension and number. However, due to the fact that basic operations are based on paillier cryptosystem with threshold decryption techniques, the maximum time required for the MA is less than 8s under the evaluation environment.

The User: As shown in Fig. 6(c), we can also learn that the computation overhead of the user is increased with the dimension and number. The reason is that, when user wants to access medical primary diagnosis service, each dimension of query vector $w_j \in w$ should be encrypted before being sent to the cloud server for privacy concerns. Thus, the computation time of the user also increased with vectors' dimension and number. However, due to the fact that basic operations are based on paillier cryptosystem with threshold decryption techniques, the maximum time required for the user is less than 500 ms under the evaluation environment.

2) Communication Overhead. In CAMPS framework, the user first deliver the query request packet $E_{Q_{U_K}} = E_{PK_{CS}}(Q_{U_K}||Sig_K)$ to the cloud server, then the cloud server will compute the intermediate result and perform partially decryption on the query, then send the packet $E_{Q_{CS}} = E_{PK_{MA}}(Q_{CS}||Sig_{CS})$ to the MA for further process. After receiving the packet $E_{Q_{CS}} = E_{PK_{MA}}(Q_{CS}||Sig_{CS})$, MA will compute the final medical diagnosis result and responds the packet $E_{R_{MA}} = E_{PK_{U_K}}(R_{MA}||Sig_{MA})$ to the user. In the real environment, we record the size of the packets, as shown in Fig. 6(d), with the increase of the dimensions and numbers of vectors in the cloud server, the communication overhead of CAMPS increases as well, when the dimension of the vectors is 11 and the number of vectors is up to 300, the total communication cost is less than 10MB under the predefined evaluation environment.

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

7. Related work

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

Skyline Computation. The skyline query has become a popular paradigm for extracting interesting objects from multi-dimensional databases. The skyline operator was first introduced to the database community by Borzsony et al. [7] with algorithm named Block Nested Loop (BNL) and Divide and Conquer (D&C). Thereafter, it was widely studied for building user's personalized queries over centralized and distributed databases. Several sequential skyline algorithms [14,26] have been designed on efficiency for centralized storage, and the Z-search algorithm proposed by Mingjie et al. [26] was the state-of-the-art skyline computation algorithm. Recently, abundant research achievements have been gained to address distributed skyline computation for big data after the first research introduced by Balke et al. [3], which supports the web information vertically partitioned into lists for extending the expressiveness of web information system. Thereafter, Park et al. [30] proposed a parallel algorithm called SKY-MR, to compute the skylines by using MapReduce. In their scheme, a Quadtree was constructed for sampling data and judging the dominance relationships among different partitions, while the cost of data IO is heavy. In order to improve the efficiency during the process of skyline queries with the MapReduce framework, Koh et al. [16] proposed two algorithms to prevent the bottleneck of centrally finding the global skyline from the local skylines by reducing the number of dominance test and performing the necessary dominance test in parallel. Zhou et al. [37] proposed an adaptive algorithm named ADSUD, which redefine the approximate global skyline probability and choose local representative tuples due to minimum probabilistic bounding rectangle adaptively. However, both centralized skyline and distributed skyline computation were well studied on improving the efficiency, while little of the works considered on the application of similarity search. Kossmann et al. [17] proposed Nearest Neighbor algorithm which used the existing R-tree nearest neighbor search to split the data space recursively, while the privacy issue was overlooked. By embedding and exploring a novel neighboring relationship among POIs, Chen et al. [8] proposed three schemes that enable efficient verification of any location-based skyline query's result returned via an untrusted service provider. Liu et al. [21] proposed a skyline computation framework across multiple domains, within the framework, a skyline result from local service providers and collaborative service providers will be securely computed to provide better services for the client with a high efficiency. In order to select the similar (or best) medical record over encrypted database, Liu et al. [20] proposed a fully secure skyline query protocol on data encrypted using semantically-secure encryption, and the new secure dominance protocol can also be used as a building block for other queries, while the overhead of computation is heavy. Moreover, Lu et al. [24] 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.

Privacy-preserving Technique. Traditional anonymization techniques such as k-anonymity [34] and l-diversity [25], 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 non-trusted server in diagnosis system obtains the medical data, it may be able to identify an individual user easily. Al-Fedaghi et al. [1] established a semi-automated methodology for measuring personal identifiable information's sensitivity starting from initial values that can be refined manually and by self-learning from previous evaluations, the experimental result shown that even seemingly benign medical information such as blood pressure can be used to identify individual

users, not to mention that ultimate information such as DNA. Organick et al. [27] propose a risk-scale system and a methodology to identify the presence of some participants in a study from DNA data, and even fully recover their DNA sequences in some cases. Hence, the anonymization techniques 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 [11], 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. Saleheen et al. [33] defined a new behavioral privacy metric based on differential privacy, then proposed a novel data substitution mechanism named mSieve with Dynamic Bayesian Network (DBN) to protect behavioral privacy. 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 [22,31], which enabled the healthcare service providers 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. Rahulamathavan et al. [31] proposed a privacy-preserving system with SVM, which can help to diagnose the user without compromising the privacy of the users and third party. Similarly, a privacy-preserving clinical diagnosis system using naive Bayesian classier was proposed by Liu et al. [22], which can also help clinician complementary to diagnose the risk of patients disease in a privacy-preserving way. Since all the encrypted operations are based on homomorphic encryption technique, 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 CAMPS framework based on a skyline diagnosis model, which has a high accuracy. Moreover, aims at the efficiency and privacy issues, the CAMPS can protect users' medical data privacy and ensure the confidentiality of diagnosis model in the untrusted cloud server. Furthermore, based on paillier cryptosystem with threshold decryption techniques, our proposed CAMPS can be easily implemented in smart terminals due to its high efficiency.

8. Conclusions

In this paper, we propose an efficient and privacy-preserving medical primary diagnosis framework (CAMPS). Within CAMPS framework, the precise diagnosis models are outsourced to the cloud server in an encrypted manner, and users can access accurate medical primary diagnosis service timely without divulging their medical data. Specifically, based on partially decryption and secure comparison techniques, a special fast secure two-party vector dominance scheme over ciphertext is proposed, with which CAMPS achieves privacy preservation of user's query and the diagnosis result, as well as the confidentiality of diagnosis models in the outsourced cloud server. Through extensive analysis, we show that CAMPS can ensure that users' medical data and healthcare service provider's diagnosis model are kept confidential, and has significantly reduce computation and communication overhead. In addition, performance evaluations via implementing CAMPS demonstrate its effectiveness in term of the real environment.

Acknowledgements

Hui Zhu is supported in part by National Key Research and Development Program of China (2017YFB0802201), National Natural Science Foundation of China (61672411, U1401251 and 81600574), Natural Science Basic Research Plan in Shaanxi Province of China (2016ZDJC-04), and China 111 Project (B16037). Ximeng Liu is supported in part by National Natural Science Foundation of China (61702105).

References

[1] S. Al-Fedaghi, A.A.R. Al-Azmi, Experimentation with personal identifiable information, Intell. Inf. Manag. 4 (04) (2012) 123.
[2] J. Andras, S. William, P. Matthias, D. Robert, Heart disease data set, 1988, (http://archive.ics.uci.edu/ml/datasets/Heart+Disease).
[3] W. Balke, U. Güntzer, J.X. Zheng, Efficient distributed skylining for web information systems, in: International Conference on Extending Database Technology, Springer, 2004, pp. 256–273.
[4] A. Belard, T. Buchman, J. Forsberg, B.K. Potter, C.J. Dente, A. Kirk, E. Elster, Precision diagnosis: a view of the clinical decision support systems (cdss) landscape through the lens of critical care, J. Clin. Monit. Comput. 31 (2) (2017) 261–271.
[5] D. Boneh, M. Franklin, Identity-based encryption from the weil pairing, in: Annual international cryptology conference, Springer, 2001, pp. 213–229.
[6] D. Boneh, B. Lynn, H. Shacham, Short signatures from the weil pairing, in: International Conference on the Theory and Application of Cryptology and Information Security, Springer, 2001, pp. 514–532.
[7] S. Borzsony, D. Kossmann, K. Stocker, The skyline operator, in: Data Engineering, 2001. Proceedings. 17th International Conference on, IEEE, 2001, pp. 421–430.
[8] W. Chen, M. Liu, R. Zhang, Y. Zhang, S. Liu, Secure outsourced skyline query processing via untrusted cloud service providers, in: INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications, IEEE, 2016, pp. 1–9.
[9] L. Constantinescu, J. Kim, D. Feng, Sparkmed: a framework for dynamic integration of multimedia medical data into distributed m-health systems, IEEE Trans. Inf. Technol. Biomed. 16 (1) (2012) 40–52.
[10] A. Donabedian, Evaluating the quality of medical care, Milbank Q. 83 (4) (2005) 691–729.
[11] C. Dwork, Differential privacy: a survey of results, in: International Conference on Theory and Applications of Models of Computation, Springer, 2008, pp. 1–19.
[12] C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in: Theory of Cryptography Conference, Springer, 2006, pp. 265–284.
[13] Y. Gao, Q. Liu, B. Zheng, L. Mou, G. Chen, Q. Li, On processing reverse k-skyband and ranked reverse skyline queries, Inf. Sci. 293 (2015) 11–34.

[14] X. Han, J. Li, D. Yang, J. Wang, Efficient skyline computation on big data, IEEE Trans. Knowl. Data Eng. 25 (11) (2013) 2521–2535.
[15] D.E. Knuth, The art of computer programming, seminumerical algorithms, 1998 2 (1998).
[16] J.-L. Koh, C.-C. Chen, C.-Y. Chan, A.L. Chen, Mapreduce skyline query processing with partitioning and distributed dominance tests, Inf. Sci. 375 (2017) 114–137.
[17] D. Kossmann, F. Ramsak, S. Rost, Shooting stars in the sky: an online algorithm for skyline queries, in: VLDB'02: Proceedings of the 28th International Conference on Very Large Databases, Elsevier, 2002, pp. 275–286.
[18] J. Lee, H. Im, G.-w. You, Optimizing skyline queries over incomplete data, Inf. Sci. 361 (2016) 14–28.
[19] H. Lin, J. Shao, C. Zhang, Y. Fang, Cam: cloud-assisted privacy preserving mobile health monitoring, IEEE Trans. Inf. Forensics Secur. 8 (6) (2013) 985–997.
[20] J. Liu, J. Yang, L. Xiong, J. Pei, Secure skyline queries on cloud platform, in: Data Engineering (ICDE), 2017 IEEE 33rd International Conference on, IEEE, 2017, pp. 633–644.
[21] X. Liu, R. Lu, J. Ma, L. Chen, H. Bao, Efficient and privacy-preserving skyline computation framework across domains, Future Gener. Comput. Syst. 62 (2016) 161–174.
[22] X. Liu, R. Lu, J. Ma, L. Chen, B. Qin, Privacy-preserving patient-centric clinical decision support system on naive bayesian classification, IEEE J. Biomed. Health Inform. 20 (2) (2016) 655–668.
[23] X. Liu, D.-N. Yang, M. Ye, W.-C. Lee, U-skyline: a new skyline query for uncertain databases, IEEE Trans. Knowl. Data Eng. 25 (4) (2013) 945–960.
[24] R. Lu, H. Zhu, X. Liu, J.K. Liu, J. Shao, Toward efficient and privacy-preserving computing in big data era, IEEE Netw. 28 (4) (2014) 46–50.
[25] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, L-Diversity: privacy beyond k-anonymity, ACM Trans. Knowl. Disc. Data (TKDD) 1 (1) (2007) 3.
[26] T. Mingjie, Y. Yu, W.G. Aref, Q. Malluhi, M. Ouzzani, Efficient parallel skyline query processing for high-dimensional data, IEEE Trans. Knowl. Data Eng. PP (99) (2018). 1–1
[27] L. Organick, S.D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M.Z. Racz, G. Kamath, P. Gopalan, B. Nguyen, et al., Random access in large-scale dna data storage, Nat. Biotechnol. 36 (3) (2018) 242.
[28] P. Paillier, Public-key cryptosystems based on composite degree residuosity classes, in: International Conference on the Theory and Applications of Cryptographic Techniques, Springer, 1999, pp. 223–238.
[29] J. Paparrizos, R.W. White, E. Horvitz, Screening for pancreatic adenocarcinoma using signals from web search logs: feasibility study and results, J. Oncol. Pract. 12 (8) (2016) 737–744.
[30] Y. Park, J.-K. Min, K. Shim, Parallel computation of skyline and reverse skyline queries using mapreduce, Proceedings of the VLDB Endowment 6 (14) (2013) 2002–2013.
[31] Y. Rahulamathavan, S. Veluru, R.C.-W. Phan, J.A. Chambers, M. Rajarajan, Privacy-preserving clinical decision support system using gaussian kernel-based classification, IEEE J. Biomed. Health Inform. 18 (1) (2014) 56–66.
[32] M. Sajid, A. Osman, G.U. Siddiqui, H.B. Kim, S.W. Kim, J.B. Ko, Y.K. Lim, K.H. Choi, All-printed highly sensitive 2d mos 2 based multi-reagent immunosensor for smartphone based point-of-care diagnosis, Sci. Rep. 7 (1) (2017) 5802.
[33] N. Saleheen, S. Chakraborty, N. Ali, M.M. Rahman, S.M. Hossain, R. Bari, E. Buder, M. Srivastava, S. Kumar, msieve: differential behavioral privacy in time series of mobile sensor data, in: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 2016, pp. 706–717.
[34] L. Sweeney, K-anonymity: a model for protecting privacy, Int. J. Uncertainty Fuzziness Knowl. Based Syst. 10 (05) (2002) 557–570.
[35] Z. Yang, Y. Huang, Y. Jiang, Y. Sun, Y.-J. Zhang, P. Luo, Clinical assistant diagnosis for electronic medical record based on convolutional neural network, Sci. Rep. 8 (1) (2018) 6329.
[36] X. Yi, A. Bouguettaya, D. Georgakopoulos, A. Song, J. Willemson, Privacy protection for wireless medical sensor data, IEEE Trans. Dependable Secure Comput. 13 (3) (2016) 369–380.
[37] X. Zhou, K. Li, Y. Zhou, K. Li, Adaptive processing for distributed skyline queries over uncertain data, IEEE Trans. Knowl. Data Eng. 28 (2) (2016) 371–384.
[38] H. Zhu, X. Liu, R. Lu, H. Li, Efficient and privacy-preserving online medical prediagnosis framework using nonlinear svm, IEEE J. Biomed. Health Inform. 21 (3) (2017) 838–850.