IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 12, 2018, accepted August 15, 2018, date of publication August 23, 2018, date of current version September 21, 2018.

*Digital Object Identifier 10.1109/ACCESS.2018.2866971*

# A Privacy-Preserving Online Medical Prediagnosis Scheme for Cloud Environment

**WEI GUO [1,2], JUN SHAO [2,3], RONGXING LU [2], (Senior Member, IEEE), YINING LIU [1], AND ALI A. GHORBANI [2], (Senior Member, IEEE)**

[1] Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China
[2] Faculty of Computer Science, Canadian Institute for Cybersecurity, University of New Brunswick, Fredericton, NB E3B 5A3, Canada
[3] School of Computer and Information Engineering, Zhejiang Gongshang University, Zhejiang 310018, China

Corresponding author: Rongxing Lu (rlu1@unb.ca)

This work was supported in part by NSERC Discovery Grants under Grant 04009, in part by NBIF Start-Up under Grant Rif 2017-012, in part by HMF under Grant HMF2017 YS-04, in part by NF under Grant NF-2017-05, in part by ZJNSF under Grant LZ18F020003, and in part by NSFC under Grants 61472364 and 61662016.

**ABSTRACT** The paradigm of online medical prediagnosis has emerged to ease the shortage of health professionals in rural areas. It can provide a 24-hour online health care service and guide rural residents' medical treatment. However, the development of online medical prediagnosis system still faces many challenges, involving the leakage and overuse of medical information. In this paper, we utilize the logistic regression to design a privacy-preserving medical prediagnosis scheme for the cloud environment, named POMP, which provides a health care service for users without violating their privacy. It is characterized by employing homomorphic encryption techniques to achieve a privacy-preserving prediagnosis process over the encrypted data. The proposed POMP scheme also adopts a preprocessing technique and Bloom filter to reduce the computational cost in the prediagnosing process. Through extensive analyses, we demonstrate that the proposed POMP scheme can resist various security threats and protect the privacy successfully. In order to evaluate the performance, we also implemented the POMP scheme and measured the running time on the smartphone and computer. The experimental result shows POMP's efficiency in terms of the computational and communication burden.

**INDEX TERMS** Homomorphic encryption, logistic regression, online medical prediagnosis, privacy preservation.

## I. INTRODUCTION

International labour organization (ILO) [1], in 2015, indicated the considerable difference in the distribution of healthcare resources between rural and urban areas worldwide, i.e., 56 per cent of people living in rural areas do not have access to the essential healthcare service—more than double the figure in urban areas where only 22 per cent are not covered. This healthcare inequity was due to the severe shortfall of health workers in rural areas, which has been a block in the way of achieving the series of public health priorities, such as reducing child and maternal mortality, increasing vaccine coverage, and battling HIV/AIDS [2]. Therefore, how to enhance the healthcare quality in rural areas has been a critical issue confronting all governments in the world.

Recently, the paradigm of online medical prediagnosis [3]–[6] has emerged and been recognized as a promising solution to the lack of health professionals in rural areas. Its core idea is combining the cloud computing and machine learning techniques for medical automation, such as the automated diagnosis and analysis, which will reduce doctors' workload and free them up for more undiagnosed patients. Therefore, for governments, the online medical prediagnosis scheme is an opportunity to improve the healthcare environment in rural areas.

Health service provider (HSP) is an essential component in the online medical prediagnosis framework. As shown in Fig. 1, the HSP is responsible for collecting a large amount of historical medical data from clinics. Using different machine learning algorithms, the HSP can train a prediagnostic model from these collected data. This model can be used to predict healthcare users' likelihood of contracting a specific disease. Subsequently, the HSP outsources the prediagnostic model to the cloud platform (CP) [7], [8], which hosts the HSP's model and offers a 24-hour online

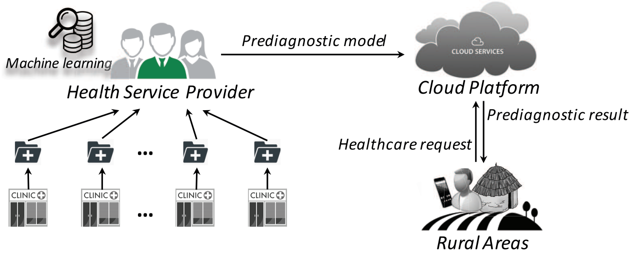


**FIGURE 1.** Framework of online medical prediagnosis.

prediagnosis service for rural areas. Based on this technique, rural residents can access the expert healthcare service at home (i.e., they use their smartphones to submit their physical symptoms and obtain the prediagnostic result).

Privacy preservation is of paramount importance to the online medical prediagnosis scheme as medical data are deeply involved in sensitive information. Generally, privacy issues can be divided into two categories based on two phases in the machine learning: the model training phase and medical prediagnosis phase. In the first phase, the HSP trains a prediagnostic model $\omega$ from historical medical records, which contain patients' sensitive information [9], [10]. Our work mainly focuses on privacy issues in second phase, where the CP examines a healthcare user's physical condition based on user's personal information. Concretely, in order to make the prediagnosis service, the CP needs to calculate the medical examination function $MedExam(\vec{x}, \omega)$, where $\vec{x}$ is a healthcare user's feature vector on her symptoms and $\omega$ is the HSP's prediagnostic model. However, the two inputs cannot be revealed directly, because the feature vector $\vec{x}$ contains the healthcare user's sensitive personal data and the prediagnostic model $\omega$ is the HSP's private asset. To preserve their privacy, our paper aims at designing a privacy-preserving online medical prediagnosis scheme, where the CP can calculate $MedExam(\vec{x}, \omega)$ but should not learn anything about the two inputs $\vec{x}$ and $\omega$.

To design this scheme, we apply the homomorphic encryption techniques [11] to protect these sensitive data. With the technique, a specific linear algebraic manipulation can be directly performed on the ciphertext. Based on this technique, many existing works [12]–[16] have been proposed to solve privacy issues in the online medical prediagnosis process. For example, Bos *et al.* [12] used a fully homomorphic cryptosystem to implement a prediagnosis scheme, which is based on the logistic regression and the Cox proportional hazard model. Particularly, we apply the BGN homomorphic cryptosystem [17] to protect the confidentiality of the user's feature vector $\vec{x}$ and the HSP's prediagnostic model $\omega$.

Over the past years, several mathematical models have been developed, studied, and used to perform the valuable prediagnosis in the healthcare, like SVM [18], Naïve Bayes [19], decision tree [20], and logistic regression (LR) [21], etc. However, the existing LR-based prediagnostic schemes [22], [23] cannot protect the confidentiality of the HSP's LR prediagnostic model. Therefore, our paper focuses on the logistic regression, fine-tunes it as a medical examination function, and then proposes a privacy-preserving online medical prediagnosis scheme. Specifically, our proposed scheme will preserve the healthcare user's privacy and the HSP's confidentiality, simultaneously. The main contributions of this paper are two-fold:

- Firstly, we propose a privacy-preserving online medical prediagnosis scheme. In this scheme, to protect the privacy, the healthcare user and the HSP encrypt the feature vector $\vec{x}$ and prediagnostic model $\omega$, respectively, and then submit them to the CP. Upon receiving these data, the CP can directly perform the medical examination function $MedExam(\cdot)$ on these encrypted data without decryption. Compared with traditional schemes, our proposed scheme not only protects the privacy of the feature vector $\vec{x}$, but also preserves the confidentiality of the prediagnostic model $\omega$.
- Secondly, the proposed POMP scheme can achieve a lightweight medical prediagnostic service by exploiting the preprocessing technique [24] and Bloom filter [25]. Furthermore, we also implement POMP over smartphone and computer,[1] and measure the running time in real environment. The experimental result demonstrates that POMP can provide an efficient medical prediagnosis service.

The remainder of this paper is organized as follows: In Section II, we introduce our system model, security requirements, and our design goals. In Section III, the logistic regression, BGN cryptosystem, and Bloom filter are recalled as the preliminaries. After that, we present our scheme in Section IV, followed by its security analysis and performance evaluation in Sections V and VI, respectively. We also discuss the related work in Section VII. Finally, we draw our conclusions in Section VIII.

## II. MODELS AND DESIGN GOALS

In this section, we describe the system model, security requirements, and identify our design goals.

### A. SYSTEM MODEL

In our system model, we mainly focus on the logistic regression to design a privacy-preserving online medical prediagnosis scheme for the cloud environment. Specifically, the system contains three kinds of entities: health service provider (HSP), healthcare users, and cloud platform (CP), as shown in Fig. 2.

- *Health Service Provider (HSP)*: We consider the HSP as an authorized organization to access historical medical data, who owns the logistic regression (LR) prediagnostic model $\omega$. In order to provide a 24-hour online service, the HSP also needs to outsource the model to the CP. However, since the model $\omega$ is a private

[1] The application and source code can be downloaded from https://www.dropbox.com/s/36evhtmvoex5zm6/AppCode.zip?dl=0

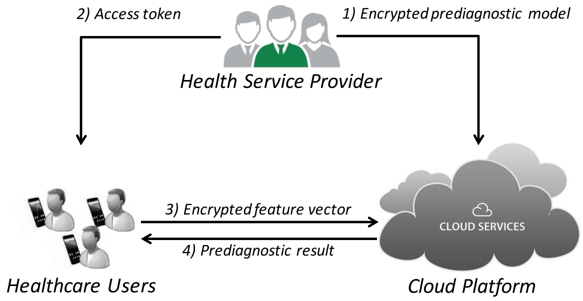


**FIGURE 2.** System model for the online medical prediagnosis.

asset, the HSP encrypts it and sends the ciphertext to the CP. Then, with this encrypted model, the CP diagnoses the healthcare user according to his feature vector $\vec{x} = \langle x_1, x_2, \ldots, x_l \rangle$, where all entities are the answers of the questionnaire *Qlist*, which contains $l$ questions $\{Q_1, Q_2, \ldots, Q_l\}$ inquiring users' physical state. Each question is two-choice, like "Do you have high blood pressure or hypertension, yes or no?". In addition, the HSP is also responsible for permitting legal users to get the online medical prediagnosis service by the access token distribution.

- *Healthcare Users*: Each healthcare user is equipped with a smartphone, which can show the questionnaire *Qlist* and communicate with the CP. At the beginning, the user completes $l$ questions in the *Qlist*, and the relevant answers make up the feature vector $\vec{x} = \langle x_1, x_2, \ldots, x_l \rangle$, where $x_i \in \{0, 1\}$ is the answer of question $Q_i$ ($x_i = 1$ means the answer is "yes", while $x_i = 0$ means the answer is "no"). Since the feature vector contains sensitive information, the user should encrypt it before submitting it to the CP.
- *Cloud Platform (CP)*: As the service agency of the HSP, the CP uses the encrypted LR prediagnostic model to evaluate whether the healthcare user has any illness. Concretely, upon receiving the request for healthcare prediagnosis, the CP performs the medical examination function *MedExam*$(\cdot)$ on the two encrypted inputs (i.e., the ciphertext of the user's feature vector $\vec{x}$ and the HSP's LR prediagnostic model $\omega$) and then returns the prediagnostic result to the user via a secure channel.

Moreover, we illustrate our POMP scheme's workflow in Fig. 2. At the beginning, the HSP outsources the encrypted LR prediagnostic model to the CP and assigns the access token to healthcare users. Then, the healthcare users request the online prediagnosis service by submitting their encrypted feature vector to the CP. Finally, without decryption, the CP can perform the prediagnosis service upon the ciphertext and return the result to the user via a secure channel.

### B. SECURITY REQUIREMENTS

The confidentiality and privacy are significant for a practical online medical prediagnosis scheme. In our security model, we consider all internal entities (HSP, healthcare user, and CP) are *honest-but-curious* [26]–[29], i.e., they would faithfully execute the operations in the protocol without launching any active attacks, but perhaps try to analyze received messages to obtain the valuable information. In addition, there exists an external adversary in the security model, who can eavesdrop on the communication channel to discover some valuable data.

In our scheme, the valuable information includes the healthcare user's feature vector and the HSP's LR prediagnostic model. To protect these data from the internal entities and external adversary, we should ensure following security requirements.

- *Confidentiality of the healthcare user's feature vector*: The feature vector contains amounts of user's sensitive information, which cannot be revealed to other internal entities or external adversary.
- *Confidentiality of the HSP's LR prediagnostic model*: With historical medical data, the HSP uses machine learning algorithms to train the LR prediagnostic model. Since the model is the intellectual property of HSP, it cannot be leaked or overheard by other internal entities or the external adversary.

Moreover, we assume that there is no collusion attack in our scheme, i.e., neither the healthcare user nor the HSP can collude with the CP to attack each other. Meanwhile, since our work mainly focuses on the privacy and confidentiality protection, other types of attack are also beyond the scope of this paper.

### C. DESIGN GOALS

In order to achieve the privacy-preserving medical prediagnosis service under the aforementioned system model and security requirements, our proposed scheme should fully guarantee the following two objectives:

- *Security requirements should be guaranteed.* As described above, if the proposed scheme does not consider the privacy of healthcare users, their highly sensitive feature vector would be disclosed to the CP or external adversary. In this case, the healthcare user would be reluctant to use this service due to the concern of privacy leakage. On the other hand, the HSP is a profit organization and its intellectual property (LR prediagnostic model) should be protected from leakage. Therefore, the proposed POMP scheme should guarantee the privacy of the healthcare user and the confidentiality of the HSP, simultaneously.
- *Computational and communicational efficiency should be achieved.* The healthcare user has limited resources in terms of computation and communication. Furthermore, although the CP is a cloud server with abundant resources, it is still challenging to guarantee its efficiency when thousands of healthcare users request the prediagnosis service at the same time. Therefore, it is necessary to ensure the computational efficiency both on the user- and CP-side.

## III. PRELIMINARIES

In this section, we review the logistic regression, BGN homomorphic cryptosystem, and Bloom filter, which will serve as the basis of our proposed scheme.

### A. LOGISTIC REGRESSION

The logistic regression is a popular tool for the binary classification, where the logistic regression (LR) classifier categorizes new coming feature vectors into two classes, such as pass/fail, win/lose, or alive/dead. In our prediagnostic cases, LR classifier divides new coming healthcare users into two groups (illness/health) by the classifier function

$$Pr(y = 1|\vec{\boldsymbol{x}}) = \frac{1}{1 + e^{-(\gamma + \vec{\boldsymbol{x}} \cdot \vec{\boldsymbol{\beta}})}}, \tag{1}$$

where $\vec{\boldsymbol{x}} = \langle x_1, x_2, \ldots, x_l \rangle$ is the healthcare user's feature vector, $\gamma$ is the intercept, and $\vec{\boldsymbol{\beta}} = \langle \beta_1, \beta_2, \ldots, \beta_l \rangle$ is the vector of regression coefficients. For convenience, Equ. (1) can also be fine-tuned into a linear function [10] by

$$\ln\left(\frac{Pr(y = 1|\vec{\boldsymbol{x}})}{1 - Pr(y = 1|\vec{\boldsymbol{x}})}\right) = \gamma + \vec{\boldsymbol{x}} \cdot \vec{\boldsymbol{\beta}}. \tag{2}$$

To decide whether the user contracts the disease, a threshold $\theta$ should be set [30]. If $\gamma + \vec{\boldsymbol{x}} \cdot \vec{\boldsymbol{\beta}} - \theta > 0$, the user is at a high chance of contracting the disease; Otherwise, the user is at a low risk.

Therefore, if we define the prediagnostic model $\omega = \{\gamma, \vec{\boldsymbol{\beta}}, \theta\}$, the fine-tuned LR classifier can be designed as the following medical examination function [30]

$$\begin{aligned} MedExam(\vec{\boldsymbol{x}}, \omega) &= \gamma + \vec{\boldsymbol{x}} \cdot \vec{\boldsymbol{\beta}} - \theta \\ &= \langle 1, x_1, \ldots, x_l, -1 \rangle \langle \gamma, \beta_1, \ldots, \beta_l, \theta \rangle, \end{aligned} \tag{3}$$

which can be represented by the product of two vectors. In this paper, we use this fine-tuned function to construct a prediagnosis service, where the healthcare user is prediagnosed with an illness if $MedExam(\vec{\boldsymbol{x}}, \omega) \geq 0$, otherwise the user is prediagnosed as a healthy one.

### B. BONEH-GOH-NISSIM (BGN) CRYPTOSYSTEM

The BGN is a kind of homomorphic cryptosystem [17], which allows some calculations on the ciphertext, like addition or multiplication. Therefore, with this property, the untrusted third party can perform some calculations on the ciphertext without access to the sensitive information. The BGN mainly contains three functions: key generation, encryption, and decryption.

1) Key generation: Given the security parameter $\tau$, a composite bilinear pairing $(N, g_0, \mathbb{G}, \mathbb{G}_T, e)$ can be created, where $N = pq$ and $p, q$ are two $\tau$-bit prime numbers, $g_0$ is a generator of group $\mathbb{G}$ with order $N$, and $e$ is bilinear mapping $\mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$. Set $h = g_0^q$, which is a generator of the subgroup of $\mathbb{G}$ with order $p$. Finally, we set the public key as $pk = (N, \mathbb{G}, \mathbb{G}_T, e, g_0, h)$ and corresponding private key as $sk = p$.
2) Encryption: We assume the message space is a set of integers $\{0, 1, \ldots, \Delta\}$ with $\Delta \ll q$. To encrypt a message $\mu$ from the set, we choose a random number $r \in Z_N$ and compute the ciphertext $C = E(\mu, r) = g_0^{\mu} h^r \in \mathbb{G}$.
3) Decryption: Given the ciphertext $C = g_0^{\mu} h^r \in \mathbb{G}$, the related message $\mu$ can be recovered by the private key $p$, i.e., calculating $C^p = (g_0^{\mu} h^r)^p = (g_0^p)^{\mu}$ and recovering $\mu$ by computing the discrete logarithm of $C^p$ to the base $g_0^p$. Since $0 \leqslant \mu \leqslant \Delta$, this discrete logarithm can be solved by Pollard's lambda algorithm [31] with computational complexity $O(\sqrt{\Delta})$.

In addition, the BGN supports following homomorphic properties:

- Multiplication from $\mathbb{G}$ to $\mathbb{G}_T$: Given two ciphertexts $C_1 = g_0^{\mu_1} h^{r_1}$ and $C_2 = g_0^{\mu_2} h^{r_2} \in \mathbb{G}$, the ciphertext $C \in \mathbb{G}_T$ of $\mu_1 \mu_2$ can be computed by
$$\begin{aligned} C &= e(C_1, C_2) = e(g_0^{\mu_1} h^{r_1}, g_0^{\mu_2} h^{r_2}) \\ &= e(g_0, g_0)^{\mu_1 \mu_2} e(g_0, h)^{\mu_1 r_2 + \mu_2 r_1 + q r_1 r_2} \end{aligned} \tag{4}$$
- Addition in $\mathbb{G}_T$ : Given two ciphertexts $C_1 = e(g_0, g_0)^{\mu_1} e(g_0, h)^{r_1}$ and $C_2 = e(g_0, g_0)^{\mu_2} e(g_0, h)^{r_2} \in \mathbb{G}_T$, we can calculate the cipher $C \in \mathbb{G}_T$ of $\mu_1 + \mu_2$ by
$$\begin{aligned} C &= C_1 C_2 \\ &= e(g_0, g_0)^{\mu_1} e(g_0, h)^{r_1} e(g_0, g_0)^{\mu_2} e(g_0, h)^{r_2} \\ &= e(g_0, g_0)^{\mu_1 + \mu_2} e(g_0, h)^{r_1 + r_2} \end{aligned} \tag{5}$$

With the aforementioned two properties, a secure vector product [32] can be designed, i.e., given two encrypted vectors $\langle C_1, C_2, \ldots, C_t \rangle$ and $\langle \bar{C}_1, \bar{C}_2, \ldots, \bar{C}_t \rangle$, where $C_i = g_0^{\mu_i} h^{r_i}$ and $\bar{C}_i = g_0^{\bar{\mu}_i} h^{\bar{r}_i}$ for $i = 1, 2., \ldots, t$, the secure product can be calculated by

$$\begin{aligned} &\langle C_1, C_2, \ldots, C_t \rangle \times \langle \bar{C}_1, \bar{C}_2, \ldots, \bar{C}_t \rangle \\ &\quad = \prod_{i=1}^{t} e(C_i, \bar{C}_i) = \prod_{i=1}^{t} e(g_0^{\mu_i} h^{r_i}, g_0^{\bar{\mu}_i} h^{\bar{r}_i}) \\ &\quad = e(g_0, g_0)^{\mu_1 \bar{\mu}_1 + \mu_2 \bar{\mu}_2 + \ldots + \mu_t \bar{\mu}_t} e(g_0, h)^{r_1 \bar{r}_1 + r_2 \bar{r}_2 + \ldots + r_t \bar{r}_t} \end{aligned} \tag{6}$$

### C. BLOOM FILTER

Bloom filter $BF(m, k)$ [33] is an efficient data structure for testing whether an element is a member of a set, where $m$ is the bit-length of the filter and $k$ is the number of hash function. Particularly, $k$ hash functions are defined as $\{H_1(\cdot), H_2(\cdot), \ldots, H_k(\cdot)\}$, each of which is defined as a mapping $H_i(\cdot) : \{0, 1\}^* \rightarrow \{0, 1, \ldots, m - 1\}$.

Bloom filter contains two functions: element addition and membership query. noitemsep

1) Element addition: In order to add an element $\alpha$ into the Bloom filter, we compute $k$ array positions $\{H_1(\alpha), H_2(\alpha), \ldots, H_k(\alpha)\}$ and then set the bits at all these positions to 1.
2) Membership query: To query whether an element $\alpha$ is a member of $BF(m, k)$, we need to check the relative $k$ positions $\{H_1(\cdot), H_2(\cdot), \ldots, H_k(\cdot)\}$. If any one bit of

these positions is 0, the element $\alpha$ definitely not in $BF(m, k)$. However, if all positions are 1, there are two distinct cases: 1) the element is contained in $BF(m, k)$; 2) these position have been set to 1 during the insertion of other elements, which is considered as the false positive [25] in the membership query.

Particularly, if $n$ elements have been added into $BF(m, k)$, the false positive probability $\mathcal{P}$ can be calculated by

$$\mathcal{P} = \left(1 - (1 - \frac{1}{m})^{kn}\right)^k \approx (1 - e^{-kn/m})^k, \tag{7}$$

which is minimized when $k = (m/n) \cdot \ln 2$ [25]. The false positive can be controlled by choosing suitable parameters of Bloom filter, as shown in Fig. 3.

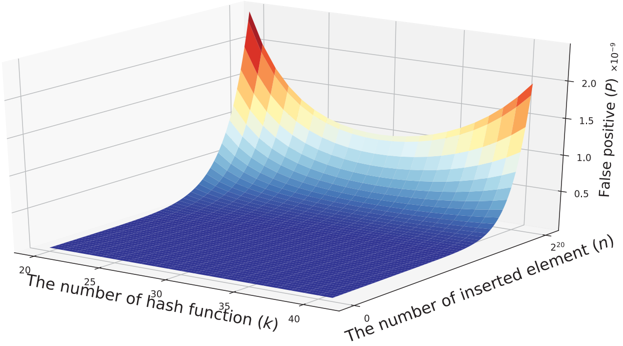


**FIGURE 3.** **False positive varying with the number of inserted element and the number of hash function.**

## IV. OUR PROPOSED SCHEME

In this section, we propose a privacy-preserving online medical prediagnosis scheme, named POMP, for the cloud environment, which mainly contains three phases: system initialization, feature vector submission, and medical prediagnosis. In the first phase, the HSP generates the public parameters for the whole system, securely outsources the LR prediagnostic model to the CP, and authorizes healthcare users by the access token. In the second phase, the healthcare user encrypts his feature vector and sends it to the CP to request the online prediagnosis service. Finally, in the third phase, the CP performs the prediagnosis service on encrypted data and returns the diagnostic result to the user in a secure manner.

### A. SYSTEM INITIALIZATION

In this phase, the HSP initializes the whole system by the following three steps: 1) the HSP generates public parameters and publishes them to other entities, 2) the HSP outsources the encrypted LR prediagnostic model to the CP, and 3) the HSP authorizes registered healthcare users by assigning the access token to them.

*Step 1 (Generating Public Parameters):* The HSP first selects the security parameter $\tau$ to generate the composite bilinear parameters $(\mathbb{G}, \mathbb{G}_T, N, g_0, e)$, where $\mathbb{G}, \mathbb{G}_T$ are two group of composite order $N = pq$, $g_0$ is a generator of group $\mathbb{G}$, and $e$ is computable mapping $\mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$. The group $\mathbb{G}$ contains two subgroups $\mathbb{G}_q$ and $\mathbb{G}_p$, which are generated by two generators $g = g_0^p$ and $h = g_0^q$, respectively. Then, the HSP calculates $\bar{g} = gh^r$, where $r$ is a random number selected from $Z_N$. Finally, the HSP publishes public parameters $(N, \mathbb{G}, \mathbb{G}_T, e, \bar{g}, h)$ to other entities and keeps other parameters $(g_0, g, r, p, q)$ secret.

*Step 2 (Outsourcing Encrypted LR Prediagnostic Model):* The original parameters of prediagnostic model are either positive/negative decimals or zeros [30]. In order to convert them into non-negative integers in $Z_N$, the HSP processes these parameters by function

$$f(x) = \lfloor 1000 * x \rfloor \ mod \ N, \tag{8}$$

i.e., expanding each parameter 1000 times and removing the decimal part. For example, a real parameter 0.0053 in [21] is converted to the integer 5 by $f(x)$. In here, we assume that all parameters of prediagnostic model $\omega = \{\gamma, \vec{\boldsymbol{\beta}}, \theta\}$ have been converted by $f(x)$.

To ensure the confidentiality, the HSP should encrypt these parameters $\gamma, \vec{\boldsymbol{\beta}} = \langle \beta_1, \beta_2, \ldots, \beta_l \rangle, \theta$ by the BGN cryptosystem

$$\begin{aligned} E_0 &= g^{\gamma} h^{r_0}, \\ E_1 &= g^{\beta_1} h^{r_1}, \quad E_2 = g^{\beta_2} h^{r_2}, \cdots E_l = g^{\beta_l} h^{r_l}, \\ E_{l+1} &= vg^{\theta} h^{r_{l+1}}, \end{aligned} \tag{9}$$

where $r_0, r_1, \ldots, r_{l+1}$ are random numbers chosen from $Z_N$. These encrypted parameters are denoted by $\Omega = \{E_0, E_1, \ldots, E_{l+1}\}$.

In addition, to ensure prediagnosis service can be performed on ciphertext, the HSP also needs to generate a Bloom filter $BF(m, k)$ containing all diseased vectors' data. It is created as follows:

1) Since the questionnaire *Qlist* contains $l$ questions and each question is two-choice, there are totally $2^l$ possible feature vectors, which compose the vector space $S$.
2) With the logistic regression, the HSP divides the space $S$ into two categories: the diseased vector space $S_d$ and the healthy vector space $S_h$. In detail, for each instance $\vec{\boldsymbol{x}} \in S$, if $MedExam(\vec{\boldsymbol{x}}, \omega) \geq 0$, $\vec{\boldsymbol{x}}$ is classified into $S_d$; otherwise $\vec{\boldsymbol{x}}$ is categorized into $S_h$.
3) For each feature vector $\vec{\boldsymbol{x}}$ in the diseased space $S_d$, the HSP calculates $e(g, g)^{MedExam(\vec{\boldsymbol{x}}, \omega)}$ and adds it into the $BF(m, k)$.

Finally, the HSP outsources his prediagnosis service to the CP by providing encrypted parameters $\Omega$ and the Bloom filter $BF(m, k)$. With these data, the CP can construct a privacy-preserving online medical prediagnosis service.

*Step 3 (Authorizing Registered User):* To give healthcare users permission to the online prediagnosis service, the HSP assigns the access token $\{A_0, A_1, \ldots, A_{l+1}\}$ to them. Each component of the access token is generated by

$$A_i = h^{r_i} \ (i = 0, 1, \ldots, l + 1), \tag{10}$$

where $r_0, r_1, \ldots, r_{l+1}$ are the same random numbers used in Eq. (9).

### B. FEATURE VECTOR SUBMISSION

The healthcare user completes the $l$ questions in $Qlist$ and generates the feature vector $\vec{\boldsymbol{x}} = \langle x_1, x_2, \ldots, x_l \rangle$, which contains sensitive data and cannot be disclosed to other entities. Therefore, to protect the privacy, the healthcare user encrypts the feature vector before submitting it to the CP. Besides, the healthcare user utilizes the access token to compute an assistant data $D$, which ensures the correctness of the medical prediagnosis.

Particularly, the healthcare user encrypts each element in the feature vector $\vec{\boldsymbol{x}} = \langle x_1, x_2, \ldots, x_l \rangle$ by

$$X_i = \bar{g}^{x_i} h^{\bar{r}_i} \ (i = 1, 2, \ldots, l), \tag{11}$$

where $\bar{r}_1, \bar{r}_2, \ldots, \bar{r}_l$ are random numbers chosen from $Z_N$. Then, the user utilizes these ciphertexts to compute the assistant data by

$$D = e(\bar{g}, A_0) \cdot \prod_{i=1}^{l} e(X_i, A_i) \cdot e(\bar{g}^{-1}, A_{l+1}), \tag{12}$$

where $A_0, A_1, \ldots, A_{l+1}$ are elements of the access token. Finally, the user requests the online prediagnosis service by submitting the encrypted feature vector $\vec{\boldsymbol{X}} = \langle X_1, X_2, \ldots, X_l \rangle$ and the assistant data $D$ to the CP.

### C. MEDICAL PREDIAGNOSIS

In this phase, the CP evaluates healthcare users' physical state by calculating the medical examination function on the ciphertexts. Then, the CP returns the prediagnostic result to healthcare users via the secure channel. The detail is described as follows:

1) Upon receiving the healthcare request from a healthcare user, the CP computes the medical examination function $MedExam(\cdot)$ with the two inputs (the encrypted feature vector $\vec{\boldsymbol{X}}$ and LR model $\Omega$) by

$$\begin{aligned} &MedExam(\vec{\boldsymbol{X}}, \Omega) \\ &\quad = e(\bar{g}, E_0) \cdot \prod_{i=1}^{l} e(X_i, E_i) \cdot e(\bar{g}^{-1}, E_{l+1}). \end{aligned} \tag{13}$$

2) Then, with the assistant data $D$ and the Bloom filter $BF(m, k)$, the CP can decide whether the healthcare user contracts a disease by computing

$$\frac{MedExam(\vec{\boldsymbol{X}}, \Omega)}{D}. \tag{14}$$

If $\frac{MedExam(\vec{\boldsymbol{X}}, \Omega)}{D} \in BF(m, k)$, the CP prediagnoses the user with the illness, otherwise the user is prediagnosed as a healthy person. Finally, via the secure channel, the CP returns the result to the user.

*Correctness:* There are two types of healthcare users, i.e., the user who is not well and the healthy user. To show the correctness of prediagnosis service, we illustrate the prediagnosis process of two typical samples, respectively.

- *Sample 1*: The first sample is a diseased user, whose feature vector $\vec{\boldsymbol{x}}_d$ belongs to the diseased vector space $S_d$. Therefore, the CP should prediagnose him with the illness. The prediagnosis process is described as follows: After receiving the encrypted feature vector $\vec{\boldsymbol{X}}$ and assistant data $D$, the CP performs the prediagnosis by computing

$$\begin{aligned} &\frac{MedExam(\vec{\boldsymbol{X}}, \Omega)}{D} \\ &\quad = \frac{e(\bar{g}, E_0) \cdot \prod_{i=1}^{l} e(X_i, E_i) \cdot e(\bar{g}^{-1}, E_{l+1})}{D} \\ &\quad \xrightarrow{E_0=g^{\gamma} h^{r_0},\ E_i=g^{\beta_i} h^{r_i} (i=1,2\ldots,l),\ E_{l+1}=g^{\theta} h^{r_{l+1}}} \\ &\quad = \frac{e(\bar{g}, g^{\gamma} h^{r_0}) \cdot \prod_{i=1}^{l} e(X_i, g^{\beta_i} h^{r_i}) \cdot e(\bar{g}^{-1}, g^{\theta} h^{r_{l+1}})}{D} \\ &\quad \xrightarrow{D=e(\bar{g},A_0)\cdot\prod_{i=1}^{l} e(X_i,A_i)\cdot e(\bar{g}^{-1},A_{l+1})\ \left(A_i=h^{r_i}(i=0,1,\ldots,l+1)\right)} \\ &\quad = \frac{e(\bar{g}, g^{\gamma} h^{r_0}) \cdot \prod_{i=1}^{l} e(X_i, g^{\beta_i} h^{r_i}) \cdot e(\bar{g}^{-1}, g^{\theta} h^{r_{l+1}})}{e(\bar{g}, h^{r_0}) \cdot \prod_{i=1}^{l} e(X_i, h^{r_i}) \cdot e(\bar{g}^{-1}, h^{r_{l+1}})} \\ &\quad = e(\bar{g}, g^{\gamma}) \cdot \prod_{i=1}^{l} e(X_i, g^{\beta_i}) \cdot e(\bar{g}^{-1}, g^{\theta}) \\ &\quad \xrightarrow{X_i=\bar{g}^{x_i} h^{\bar{r}_i},\ \bar{g}=gh^{r},\ e(h,g)=1} \\ &\quad = e(g, g^{\gamma}) \cdot \prod_{i=1}^{l} e(g^{x_i}, g^{\beta_i}) \cdot e(g^{-1}, g^{\theta}) \\ &\quad = e(g, g)^{\gamma+\sum_{i=1}^{l} x_i \beta_i - \theta} = e(g, g)^{MedExam(\vec{\boldsymbol{x}}_d, \omega)} \end{aligned}$$

Since this user's feature vector $\vec{\boldsymbol{x}}_d$ belongs to the diseased space $S_d$ and $e(g, g)^{MedExam(\vec{\boldsymbol{x}}_d, \omega)}$ has been added into $BF(m, k)$, the CP diagnoses this user with the illness.

- *Sample 2*: The second sample is a healthy user, whose feature vector $\vec{\boldsymbol{x}}_h$ belongs to the healthy vector space $S_h$. Therefore, the CP should prediagnose this user as a healthy person. The prediagnosis process is described as follows: the CP also calculates the $\frac{MedExam(\vec{\boldsymbol{X}}, \Omega)}{D}$ and then checks whether it has been contained in $BF(m, k)$. The second sample's feature vector $\vec{\boldsymbol{x}}_h$ belongs to the healthy vector space $S_h$ and $e(g, g)^{MedExam(\vec{\boldsymbol{x}}_h, \omega)}$ has not been included into $BF(m, k)$. Therefore, the CP should return a healthy result to the healthcare user. However, due to the false positive of Bloom filter, the CP may misdiagnose the healthy user with the false positive $\mathcal{P}$. To ensure the prediagnostic correctness, we can control the false positive by choosing suitable parameters $m$ and $k$ of Bloom filter. According to [34], given the expected false positive $\mathcal{P}$ and the maximum number of inserted element $n$, the bit-length of Bloom filter $m$ should be

$$m = -\frac{n \cdot \ln \mathcal{P}}{(\ln 2)^2}. \tag{15}$$

For example, when $\mathcal{P} = 10^{-9}$ and $n = 2^{20}$, we can calculate the bit-length $m$ of Bloom filter is $45, 227, 980$ bit$\approx$ 5.39 MB. In this situation, we also measure the false positive varying with the number of hash function $k$, and the number of inserted element $n$, as shown in Fig. 3. When the number $k$ of hash function equals 30, the false positive can be controlled less than $10^{-9}$. Therefore, it is possible to control the false positive $\mathcal{P}$ to ensure the correctness of the prediagnosis service.

## V. SECURITY ANALYSIS

In this section, we analyze the security properties of the proposed POMP scheme. In particular, following the aforementioned security requirements, our analysis will focus on how the proposed scheme can achieve the confidentiality of LR parameters and feature vector.

### A. CONFIDENTIALITY OF HSP'S LR PREDIAGNOSTIC MODEL

In order to preserve the confidentiality, the HSP encrypts all parameters of the LR prediagnostic model before outsourcing them to the CP. In the following analysis, we illustrate that HSP's confidential data can be preserved from the external adversary, healthcare user, and CP.

- *HSP's confidential data is preserved from the external adversary.* In the proposed scheme, LR parameters $\omega = \{\gamma, \beta_1, \ldots, \beta_l, \theta\}$ owned by the HSP are encrypted into the ciphertext $\Omega = \{E_0, E_1, \ldots, E_l, E_{l+1}\}$ by the BGN cryptosystem:

$$\begin{aligned} E_0 &= g^{\gamma} h^{r_0}, \\ E_1 &= g^{\beta_1} h^{r_1}, \ldots, E_l = g^{\beta_l} h^{r_l}, \\ E_{l+1} &= g^{\theta} h^{r_{l+1}}. \end{aligned} \tag{16}$$

  Since the BGN is semantic secure against the chosen plaintext attack, the parameters $\{\gamma, \beta_1, \ldots, \beta_l, \theta\}$ are also semantic secure and privacy-preserving. Therefore, even thought the external adversary eavesdrops these ciphertexts $\{E_0, E_1, \ldots, E_l, E_{l+1}\}$, he still cannot identify the corresponding contents.
- *HSP's confidential data is protected from the healthcare user.* In our system model, the healthcare user is considered as an *honest-but-curious* user, i.e., he would not actively attack the HSP but try to infer its confidential information from the received data. However, the healthcare user only obtains the access token from the HSP, which does not involve any information on LR parameters. Therefore, the HSP's confidential data is also protected from the healthcare user.
- *HSP's confidential data is guarded from the CP.* In the proposed scheme, the HSP outsources the encrypted LR model $\Omega$ to the CP. As an *honest-but-curious* entity, the CP attempts to infer HSP's confidential information from the received data. In our security model, we assume that there is no collusion attack. However, to show the security, we demonstrate that POMP can still protect the HSP's confidential data from the CP even when he colludes with the healthcare user. For instance, in order to recover $\gamma - \theta$, the CP asks the healthcare user to generate a zero vector $\vec{\boldsymbol{x}} = \langle 0, 0, \ldots, 0 \rangle$ and submit the encrypted zero vectors $\vec{\boldsymbol{X}} = \langle X_1, X_2, \ldots, X_l \rangle$ where

$$X_1 = h^{\bar{r}_1}, X_2 = h^{\bar{r}_2}, \ldots, X_l = h^{\bar{r}_l}. \tag{17}$$

  Meanwhile, the healthcare user generates and submits the assistant data $D$ to the CP, where

$$D = e(h, h)^{rr_0 + \sum_{i=1}^{l} \bar{r}_i r_i - r r_{l+1}}. \tag{18}$$

  Upon receiving $\vec{\boldsymbol{X}}$ and $D$, the CP can calculate the examination function $MedExam(\vec{\boldsymbol{X}}, \Omega)$ and then compute

$$\frac{MedExam(\vec{\boldsymbol{X}}, \Omega)}{D} = e(g, g)^{\gamma - \theta}. \tag{19}$$

  However, since the element $g$ is the secret data of the HSP, the CP cannot know the base $e(g, g)$ and also cannot retrieve $\gamma - \theta$ from the result. Therefore, even though cooperating with the healthcare user, the CP still cannot violate the HSP's confidential data.

### B. CONFIDENTIALITY OF HEALTHCARE USER'S FEATURE VECTOR

The healthcare user also uses the BGN cryptosystem to preserve the feature vector $\vec{\boldsymbol{x}} = \langle x_1, x_2, \ldots, x_l \rangle$ from the CP, external adversary, and HSP. Particularly, healthcare user encrypts the feature vector by

$$X_i = \bar{g}^{x_i} h^{\bar{r}_i} \ (i = 1, 2, \ldots, l). \tag{20}$$

Then, the user requests the online prediagnosis service by submitting the encrypted feature vector to the CP. Particularly, for the CP and external adversary, they can receive or eavesdrop the encrypted feature vector, but cannot recover the plaintext from these ciphertexts due to the lack of the private key $p$. For the HSP, he holds the private key $p$, but cannot get the encrypted feature vector. Therefore, the HSP also cannot violate the user's privacy. In summary, the healthcare user's privacy is protected from the CP, external adversary, and HSP.

## VI. PERFORMANCE EVALUATION

In this section, we evaluate the performance of the proposed POMP scheme in terms of the computational costs and communication overheads. Particularly, we first theoretically analyze its computational and communication complexity, and then perform an experimental evaluation in the real environment.

### A. THEORETICAL ANALYSIS

We theoretically analyze the computation complexity of the HSP, healthcare user, and CP, respectively, and then discuss the communication overheads between the healthcare user and CP.

#### 1) COMPUTATIONAL COSTS

The computational cost of our proposed POMP scheme is mainly affected by the following five time-consuming operations: the exponentiation in $\mathbb{G}$, the multiplication in $\mathbb{G}$, the multiplication in $\mathbb{G}_T$, the pairing operation, and the hash function, denoted by $C_e$, $C_{m1}$, $C_{m2}$, $C_p$ and $C_h$, respectively. Based on these five operations, we analyze each entity's computational complexity.

- *Computational complexity of the HSP:* In the system initialization, the HSP outsources the encrypted LR parameters and the Bloom filter to the CP, and assigns the access token to the healthcare user. Particularly, the HSP takes $2(l+2) * C_e + (l+2) * C_{m1}$ to encrypt the LR parameters, $2^l * (C_e + C_p)$ to initialize the Bloom filter, and $(l+2) * C_e$ to calculate the access token. These calculations can be computed offline and will not affect the daily request/feedback.
- *Computational complexity of the healthcare user:* The healthcare user launches the prediagnosis service by submitting the encrypted feature vector and the assistant data to the CP. Especially, the user takes $2l * C_e + l * C_{m1}$ to encrypt his feature vector and $(l+2)*C_p+(l+1)*C_{m2}$ to compute the assistant data.
- *Computational complexity of the CP:* The CP performs the prediagnosis service on the ciphertext. Specifically, the CP requires $(l+2) * C_p + (l+2) * C_{m2}$ to compute the prediagnosis function and $k * C_h$ for the membership query in Bloom filter.

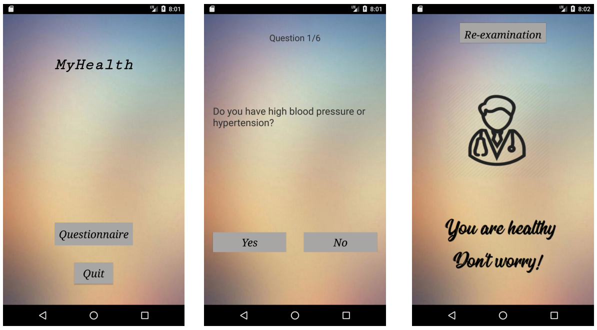
#### 2) COMMUNICATION OVERHEADS

The communications of the proposed POMP scheme can be divided into three parts: HSP-to-CP, HSP-to-user, and user-to-CP. The first two parts only happen once in the system initialization phase, which will not influence the communication efficiency of the whole system.

Therefore, we mainly focus on the user-to-CP communication, where the healthcare user generates and sends their request to the CP. The format of this request is $X_1||X_2||\dots||X_l||D$, where "||" is the concatenation operation. Particularly, if we assume the length of BGN cipher is 2048-bit, the whole size of data should be $(l+1)*2,048$ bits. Obviously, this communication overhead is linear with the number of questions in the questionnaire. Therefore, the proposed POMP scheme has an accaptable communication and is suitable for the real environment.

### B. EXPERIMENTAL EVALUATION

In our experiment, we implemented the proposed POMP scheme using the Java and the JPBC library [24], and then measured the running time in real environment.



Do you have high blood pressure or hypertension?

You are healthy Don't worry!

**FIGURE 4.** Android application to simulate healthcare user.

As shown in Fig. 4, we developed an Android application for the healthcare user, which runs on a smartphone with 2.30 GHz processor, 4 GB memory, and Android 7.0 system. We also developed two applications for the HSP and CP, which are deployed on a computer with the Intel i5-2450M, 2.50 GHz processor, and 8 GB memory.

#### 1) EXPERIMENTAL EVALUATION OF THE HSP

In our scheme, the HSP is responsible for initializing the whole system by generating the Bloom filter, the encrypted LR prediagnostic model and the access token. Particularly, in the real environment, the initial time of Bloom filter is recorded in TABLE 1. It shows that the computational cost varies with the number of questions $l$ from 6 to 20 with the increment of 2. In addition, we also evaluate the running time of the LR parameters encryption and the access token generation, which are depicted in Fig. 5. Obviously, both of their running time are linear with $l$ and larger than 5,000 ms when $l = 20$. However, these computational costs are acceptable, since these operations are computed only once and will not affect the efficiency of the online prediagnosis service.

**TABLE 1.** HSP's running time for initializing Bloom filter.

| $l$ | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 |
|---|---|---|---|---|---|---|---|---|
| Time (ms) | 8 | 88 | 210 | 614 | 3,396 | 6,822 | 22,674 | 92,661 |

#### 2) EXPERIMENTAL EVALUATION OF THE HEALTHCARE USER

The healthcare user utilizes the pairing operation to compute the assistant data. However, since the pairing operation is time-consuming, it is unsuitable for being deployed on the smartphone. To solve this problem, we use the preprocessing technique [24] to accelerate the calculation speed of generating assistant data. Its core idea is to pre-compute and store some values that will be used several times in the future. Particularly, the healthcare user calculates and stores $e(\bar{g}, A_i)$ (for $i = 0, 1, \dots, l+1$) and $e(h, A_i)$ (for $i = 1, 2, \dots, l$) in advance. Then, when computing the assistant data, the healthcare user reads these data directly from the

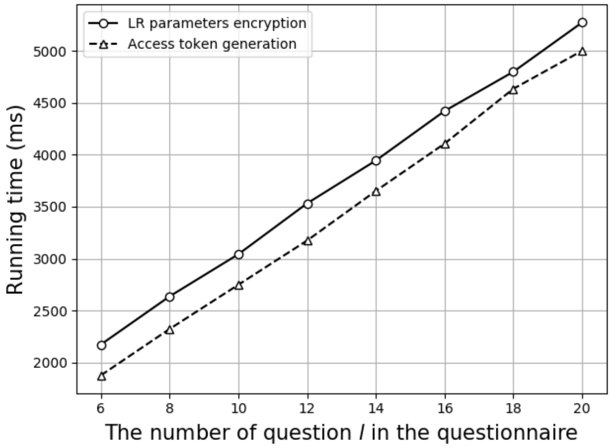


**FIGURE 5. HSP's running time of the LR parameters encryption and access token generation.**

storage and calculates

$$D = e(\bar{g}, A_0) \cdot \prod_{i=1}^{l} \left( e(\bar{g}, A_i)^{x_i} e(h, A_i)^{\bar{r}_i} \right) \cdot e(\bar{g}, A_{l+1})^{-1}. \quad (21)$$

This will accelerate the calculation of healthcare user by reducing the complexity to $2l * C_e + (2l+1) * C_{m2}$. As shown in Fig. 6, we measure the running time of the accelerated version and meanwhile compare it with the original, respectively. The result shows the preprocessing technique dramatically reduces the complexity and makes our scheme efficient for the smartphone.

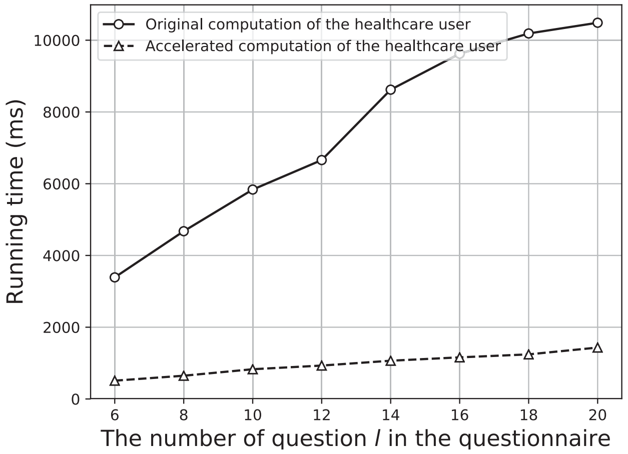


**FIGURE 6. Healthcare user's running time of the original and accelerated computation.**

### 3) EXPERIMENTAL EVALUATION OF THE CP

The CP's calculation can also be accelerated by the preprocessing technique, i.e., computing and storing $e(\bar{g}, E_0)$ and $e(\bar{g}, E_{l+1})^{-1}$ in advance, and then reading them directly when calculating the

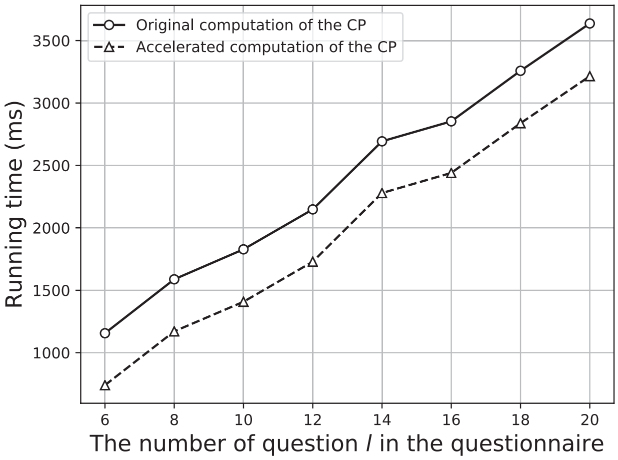


**FIGURE 7. CP's running time of the original and accelerated computation.**

$$MedExam(\vec{\boldsymbol{X}}, \Omega) = e(\bar{g}, E_0) \cdot \prod_{i=1}^{l} e(X_i, E_i) \cdot e(\bar{g}, E_{l+1})^{-1}. \quad (22)$$

By this method, the whole complexity of the CP can be reduced to $l * C_p + (l+2) * C_{m2} + k * C_h$. As shown in Fig. 7, the running time of original and accelerated CP rises steadily with the growth of $l$. Both of their running time are larger than $3,000$ ms when $l$ grows to 20. However, since the CP is a powerful cloud server, it is acceptable for it to address this computational burden.

In summary, our analyses prove that the proposed scheme is efficient in terms of computational costs and communicational overheads, which are suitable for the real environment.

## VII. RELATED WORK

Existing works on the privacy-preserving medical prediagnosis can be generally divided into two categories: the privacy-preserving training process on the historical medical data and the privacy-preserving medical prediagnosis to the new coming healthcare user. The aim of the first category is to privately train a prediagnostic model upon protected medical records. However, our work mainly focuses on the second category, which protects the healthcare user's privacy and HSP's confidentiality in the prediagnosis process.

### A. PRIVACY-PRESERVING TRAINING PROCESS ON THE HISTORICAL MEDICAL DATA

In the training process, the machine learning algorithms directly operate on the plaintext of historical medical records, which would disclose patients' sensitive data. To solve this problem, many methods have been proposed to preserve the privacy in this process.

The randomization is a prospective tool to guard the privacy in the training process. Several schemes [35]–[37] preserved the medical data by the random rotation perturbation, random matrix, or random noise [38]. These schemes can be directly performed on the randomized data. Subsequently, the correct trained model can be recovered by the de-randomization process.

The cryptographic method is another useful tool to solve this problem. For instance, some schemes [39]–[41] ensured the training algorithms can be performed upon encrypted medical records. Based on these works, Lauter *et al.* [42] presented privacy-preserving genomic algorithms over the ciphertext, which was encrypted by the fully homomorphic encryption.

### B. PRIVACY-PRESERVING MEDICAL PREDIAGNOSIS TO THE HEALTHCARE USER

In the prediagnosis service, a healthcare user submits their feature vector, which contains their physical information. To preserve the healthcare user's privacy, many schemes have been proposed recently.

In 2014, using the fully homomorphic encryption, Bos *et al.* [12] implemented a cloud server for performing private prediagnosis service upon the encrypted medical data. In their setting, the patient's medical data can be well protected by a lattice-based homomorphic cryptosystem [43]. However, in Bos *et al.*'s scheme, the confidentiality of prediagnostic model is ignored, i.e., the model is considered as the public information known by everyone (including patients). Therefore, Bos *et al.*'s scheme only achieves a privacy-preserving prediagnosis in a weaker security setting.

Bost *et al.* [13] achieved a high-level security, i.e., both the confidentiality of prediagnostic model and the privacy of feature vector are preserved. Particularly, the prediagnostic model is offered by a service provider, while the feature vector is provided by the healthcare user. Both parties need to keep their data private. To achieve this objective, Bost *et al.* [13] used the Paillier cryptosystem [44] to design three major privacy-preserving classifiers (hyperplane decision, Naïve Bayes, and decision trees) as medical examination function. Subsequently, in order to improve the efficiency, Wu *et al.* [45] employed a novel conditional oblivious transfer protocol to design an efficient privacy-preserving classifier. Jia *et al.* [46] applied the oblivious evaluation of multivariate polynomials [47] and the oblivious transfer protocol to achieve a privacy-preserving SVM classifier. Without using any time-consuming homomorphic encryptions, Jia *et al.*'s scheme achieved the efficiency in terms of communication and computation. Recently, Zhu *et al.* [15] exploited the lightweight multiparty random masking and polynomial aggregation techniques to design a medical prediagnosis framework, which is based on the nonlinear kernel SVM. In Zhu *et al.*'s scheme, both the privacy of user's feature vector and the confidentiality of SVM classifier are protected, and meanwhile it has lower overhead than [48] in terms of computation and communication.

Different from all of the aforementioned works, our proposed POMP scheme achieves a privacy-preserving prediagnosis service by using the logistic regression, which can be widely applied in the medical examination. In particular, POMP not only preserves the privacy of healthcare user's feature vector, but also protects the confidentiality of HSP's prediagnosis mode.

## VIII. CONCLUSIONS

In this paper, we have proposed an efficient and privacy-preserving online medical prediagnosis scheme for cloud environment by the BGN homomorphic cryptosystem. In our scheme, the online prediagnosis service can directly be performed on the ciphertext. Compared with the traditional medical prediagnosis based on the LR classifier, our scheme can achieve a high-level privacy setting, which protects the privacy of personal feature vector and sensitive LR parameters. To improve the efficiency, we utilized the preprocessing technique and the Bloom filter to accelerate the prediagnosis process. Then, we carried out the security analysis to demonstrate the security strength and privacy-preserving ability of the proposed scheme. Finally, the performance evaluation in real environment also showed our scheme's efficiency in terms of the computational and communication burden.

## REFERENCES

[1] X. Scheil-Adlung, "Global evidence on inequities in rural health protection. New data on rural deficits in health coverage for 174 countries," Dept. Social Protection, Int. Labour Org., Geneva, Switzerland, Tech. Rep. 47, 2015. [Online]. Available: http://www.ilo.org/wcmsp5/groups/public/-ed_protect/soc_sec/documents/publication/wcms_383890.pdf

[2] B. M. Kuehn, "Global shortage of health workers, brain drain stress developing countries," *JAMA*, vol. 298, no. 16, pp. 1853–1855, 2007.

[3] 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.

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

[5] X. Liu, R. Deng, K.-K. R. Choo, and Y. Yang, "Privacy-preserving outsourced clinical decision support system in the cloud," *IEEE Trans. Services Comput.*, to be published.

[6] C. Zhang, L. Zhu, C. Xu, and R. Lu, "PPDP: An efficient and privacy-preserving disease prediction scheme in cloud-based e-healthcare system," *Future Gener. Comput. Syst.*, vol. 79, pp. 16–25, Feb. 2018.

[7] C. Zuo, J. Shao, G. Wei, M. Xie, and M. Ji, "CCA-secure ABE with outsourced decryption for fog computing," *Future Generat. Comput. Syst.*, vol. 78, pp. 730–738, Jan. 2018.

[8] C. Zuo, J. Shao, J. K. Liu, G. Wei, and Y. Ling, "Fine-grained two-factor protection mechanism for data sharing in cloud storage," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 1, pp. 186–196, Jan. 2018.

[9] H. Kikuchi, H. Yasunaga, H. Matsui, and C.-I. Fan, "Efficient privacy-preserving logistic regression with iteratively Re-weighted least squares," in *Proc. 11th Asia Joint Conf. Inf. Secur. (AsiaJCIS)*, Fukuoka, Japan, Aug. 2016, pp. 48–54.

[10] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Statist. Softw.*, vol. 33, no. 1, pp. 1–22, 2010.

[11] Y. Zhu, Z. Huang, and T. Takagi, "Secure and controllable $k$-NN query over encrypted cloud data with key confidentiality," *J. Parallel Distrib. Comput.*, vol. 89, pp. 1–12, Mar. 2016.

[12] J. W. Bos, K. Lauter, and M. Naehrig, "Private predictive analysis on encrypted medical data," *J. Biomed. Inform.*, vol. 50, pp. 234–243, Aug. 2014.

[13] R. Bost, R. A. Popa, S. Tu, and S. Goldwasser, "Machine learning classification over encrypted data," in *Proc. 22nd Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, San Diego, CA, USA, Feb. 2015, pp. 1–14.

[14] Y. Aono, T. Hayashi, L. T. Phong, and L. Wang, "Privacy-preserving logistic regression with distributed data sources via homomorphic encryption," *IEICE Trans. Inf. Syst.*, vol. 99.D, no. 8, pp. 2079–2089, 2016.

[15] 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.

[16] J. Hua *et al.*, "CINEMA: Efficient and privacy-preserving online medical primary diagnosis with skyline query," *IEEE Internet Things J.*, to be published, doi: 10.1109/JIOT.2018.2834156.

[17] D. Boneh, E.-J. Goh, and K. Nissim, "Evaluating 2-DNF formulas on ciphertexts," in *Proc. Theory Cryptogr. Conf. (TCC)*, Cambridge, MA, USA, Feb. 2005, pp. 325–341.

[18] K. Polat, S. Güneş, and A. Arslan, "A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 482–487, 2008.

[19] Y. Wang *et al.*, "Gene selection from microarray data for cancer classification—A machine learning approach," *Comput. Biol. Chem.*, vol. 29, no. 1, pp. 37–46, 2005.

[20] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer Inform.*, vol. 2, pp. 59–78, Jan. 2006.

[21] T. J. Cleophas and A. H. Zwinderman, "Logistic regression for health profiling," in *Machine Learning in Medicine*. Dordrecht, The Netherlands: Springer, 2013, pp. 17–24.

[22] S. Carpov, T. H. Nguyen, R. Sirdey, G. Constantino, and F. Martinelli, "Practical privacy-preserving medical diagnosis using homomorphic encryption," in *Proc. 9th IEEE Int. Conf. Cloud Comput. (CLOUD)*, San Francisco, CA, USA, Jun./Jul. 2016, pp. 593–599.

[23] X. D. Zhu, H. Li, and F. H. Li, "Privacy-preserving logistic regression outsourcing in cloud computing," *Int. J. Grid Utility Comput.*, vol. 4, nos. 2–3, pp. 144–150, Sep. 2013.

[24] A. De Caro and V. Iovino, "JPBC: Java pairing based cryptography," in *Proc. 16th IEEE Symp. Comput. Commun. (ISCC)*, Kerkyra, Greece, Jun./Jul. 2011, pp. 850–855.

[25] A. Broder and M. Mitzenmacher, "Network applications of bloom filters: A survey," *Internet Math.*, vol. 1, no. 4, pp. 485–509, 2003.

[26] G. Liu, Z. Yan, and A. W. Pedryczc, "Data collection for attack detection and security measurement in mobile ad hoc networks: A survey," *J. Netw. Comput. Appl.*, vol. 105, pp. 105–122, Mar. 2018.

[27] J. Shao and G. Wei, "Secure outsourced computation in connected vehicular cloud computing," *IEEE Netw.*, vol. 32, no. 3, pp. 36–41, May/Jun. 2018.

[28] Y. Liu, W. Guo, C.-I. Fan, L. Chang, and C. Cheng, "A practical privacy-preserving data aggregation (3PDA) scheme for smart grid," *IEEE Trans. Ind. Informat.*, to be published, doi: 10.1109/TII.2018.2809672.

[29] Y. Liu, Q. Zhong, L. Chang, Z. Xia, D. He, and C. Cheng, "A secure data backup scheme using multi-factor authentication," *IET Inf. Secur.*, vol. 11, no. 5, pp. 250–255, 2017.

[30] G. Wang, R. Lu, and C. Huang, "PGuide: An efficient and privacy-preserving smartphone-based pre-clinical guidance scheme," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.

[31] A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone, *Handbook of Applied Cryptography*. Boca Raton, FL, USA: CRC Press, 1996.

[32] J. Yuan and S. Yu, "Privacy preserving back-propagation neural network learning made practical with cloud computing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 1, pp. 212–221, Jan. 2014.

[33] D. Yang, D. Tian, J. Gong, S. Gao, T. Yang, and X. Li, "Difference bloom filter: A probabilistic structure for multi-set membership query," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.

[34] S. Tarkoma, C. E. Rothenberg, and E. Lagerspetz, "Theory and practice of bloom filters for distributed systems," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 1, pp. 131–155, 1st Quart.,2012.

[35] K. Chen and L. Liu, "Privacy preserving data classification with rotation perturbation," in *Proc. 5th IEEE Int. Conf. Data Mining (ICDM)*, Houston, TX, USA, Nov. 2005, pp. 589–592.

[36] O. L. Mangasarian and E. W. Wild, "Privacy-preserving classification of horizontally partitioned data via random kernels," in *Proc. Int. Conf. Data Mining (DMIN)*, Las Vegas, NV, USA, Jul. 2008, pp. 473–479.

[37] K. Chaudhuri and C. Monteleoni, "Privacy-preserving logistic regression," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 289–296.

[38] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Theory Cryptogr. Conf.* Berlin, Germany: Springer, 2006, pp. 265–284.

[39] S. Laur, H. Lipmaa, and T. Mielikäinen, "Cryptographically private support vector machines," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 618–624.

[40] C. Orlandi, A. Piva, and M. Barni, "Oblivious neural network computing via homomorphic encryption," *EURASIP J. Inf. Secur.*, vol. 2007, p. 037343, Dec. 2007.

[41] J. Vaidya, H. Yu, and X. Jiang, "Privacy-preserving SVM classification," *Knowl. Inf. Syst.*, vol. 14, no. 2, pp. 161–178, Feb. 2008.

[42] K. Lauter, A. López-Alt, and M. Naehrig, "Private computation on encrypted genomic data," in *Proc. Int. Conf. Cryptol. Inf. Secur. Latin Amer.* Cham, Switzerland: Springer, 2014, pp. 3–27.

[43] J. W. Bos, K. Lauter, J. Loftus, and M. Naehrig, "Improved security for a ring-based fully homomorphic encryption scheme," in *Proc. IMA Int. Conf. Cryptogr. Coding*. Berlin, Germany: Springer, 2013, pp. 45–64.

[44] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *Proc. Int. Conf. Theory Appl. Cryptograph. Techn. Adv. Cryptol. (EUROCRYPT)*, Prague, Czech Republic, May 1999, pp. 223–238.

[45] D. J. Wu, T. Feng, M. Naehrig, and K. Lauter, "Privately evaluating decision trees and random forests," in *Proc. PoPETs*, vol. 4, 2016, pp. 335–355.

[46] Q. Jia, L. Guo, Z. Jin, and Y. Fang, "Privacy-preserving data classification and similarity evaluation for distributed systems," in *Proc. IEEE 36th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jun. 2016, pp. 690–699.

[47] T. Tassa, A. Jarrous, and Y. Ben-Ya'akov, "Oblivious evaluation of multivariate polynomials," *J. Math. Cryptol.*, vol. 7, no. 1, pp. 1–29, 2013.

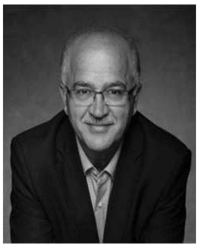
[48] Y. Rahulamathavan, S. Veluru, R. C.-W. Phan, J. A. Chambers, and M. Rajarajan, "Privacy-preserving clinical decision support system using Gaussian kernel-based classification," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 1, pp. 56–66, Jan. 2014.

**WEI GUO** received the B.S. degree in information and computational science from the Guilin University of Electronic Technology, Guilin, China, in 2015, where he is currently pursuing the master's degree with the School of Computer and Information Security. His research interests focus on applied cryptography.

**JUN SHAO** received the Ph.D. degree from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2008. He held a post-doctoral position with the School of Information Sciences and Technology, The Pennsylvania State University, USA, from 2008 to 2010. He is currently a Full Professor with the School of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou, China. His research interests include network security and applied cryptography.

**RONGXING LU** (S'09–M'10–SM'15) received the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Waterloo, Canada, in 2012. He was a Post-Doctoral Fellow with the University of Waterloo from 2012 to 2013. He was an Assistant Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, from 2013 to 2016. He has been an Assistant Professor with the Faculty of Computer Science, University of New Brunswick, Canada, since 2016. His research interests include applied cryptography, privacy enhancing technologies, and Internet of Things big data security and privacy. He is currently a Senior Member of the IEEE Communications Society. He was a recipient of the most prestigious Governor General's Gold Medal and the 8th IEEE Communications Society (ComSoc) Asia–Pacific Outstanding Young Researcher Award in 2013. He currently serves as a Vice Chair (Publication) for the IEEE ComSoc CIS-TC.

**YINING LIU** received the B.S. degree in applied mathematics from Information Engineering University, Zhengzhou, China, in 1995, the M.S. degree in computer software and theory from the Huazhong University of Science and Technology, Wuhan, China, in 2003, and the Ph.D. degree in mathematics from Hubei University, Wuhan, in 2007. He is currently a Professor with the School of Computer and Information Security, Guilin University of Electronic Technology, Guilin, China. His research interests include the information security protocol and data privacy.



**ALI A. GHORBANI** (SM'94) held variety of positions in academia for the past 35 years. Since 2008, he has been the Dean of the Faculty of Computer Science, University of New Brunswick, where he is currently the Director of the Canadian Institute for Cybersecurity. He is the Canada Research Chair (Tier 1) in Cybersecurity. He has supervised over 160 research associates, post-doctoral fellows, graduate students, and undergraduate students during his career. He has developed a number of technologies that have been adopted by high-tech companies. He co-founded two startups, Sentrant and EyesOver in 2013 and 2015. He has authored over 200 peer-reviewed articles during his career. He has authored the book *Intrusion Detection and Prevention Systems: Concepts and Techniques* published by (Springer, 2010). He holds three co-invented awarded patents in the area of network security and Web intelligence. He was a recipient of the University of New Brunswick's Research Scholar Award in 2007. He was twice one of the three finalists for the Special Recognition Award at the 2013 and 2016 New Brunswick KIRA Award for the knowledge industry. Since 2010, he has obtained more than U.S. $10M to fund six large multi-project research initiatives. He is the Co-Editor-In-Chief of the *Computational Intelligence Journal*.