IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 20, 2019, accepted November 6, 2019, date of publication November 14, 2019,
date of current version November 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2953495

Privacy-Preserving Collaborative Model Learning Scheme for E-Healthcare

FENGWEI WANG[1], (Student Member, IEEE), HUI ZHU[1], (Senior Member, IEEE), XIMENG LIU[2], (Member, IEEE), RONGXING LU[3], (Senior Member, IEEE), JIAFENG HUA[1], (Student Member, IEEE), HUI LI[1], (Member, IEEE), AND HAO LI[4]

[1]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China
[2]College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350002, China
[3]Faculty of Computer Science, University of New Brunswick, Fredericton, NB E3B 5A3, Canada
[4]The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, China

Corresponding author: Hui Zhu (zhuhui@xidian.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0802200, in part by the National Natural Science Foundation of China under Grant 61972304, Grant 61672411, and Grant 61932015, in part by the National Natural Science Foundation of Shaanxi Province under Grant 2019ZDLGY12-02, in part by the Shaanxi Innovation Team Project under Grant 2018TD-007, in part by the China Scholarship Council, Fundamental Research Funds for the Central Universities, and in part by the Innovation Fund of Xidian University.

ABSTRACT With the advances of data mining and the pervasiveness of cloud computing, online medical diagnosis service has been extensively applied in e-heathcare field, and brought great conveniences to people's life. However, due to the insufficient data sharing among healthcare centers under the security and privacy concerns of medical information, the flourish of online medical diagnosis service still faces many severe challenges including diagnostic accuracy issues. In this paper, in order to address the security issues and improve the accuracy of online medical diagnosis service, we propose a new privacy-preserving collaborative model learning scheme with skyline computation, called PCML. With PCML, healthcare centers can securely learn a global diagnosis model with their local diagnosis models in the assistance of cloud, and the sensitive medical data of each healthcare center is well protected. Specifically, with a secure multi-party vector comparison algorithm (SMVC), all local diagnosis models are encrypted by their owners before being sent to the cloud, and can be directly operated without decryption. Detailed security analysis shows that PCML can resist security threats in the semi-honest model. Moreover, PCML is implemented with medical datasets from UCI machine learning repository, and extensive simulation results demonstrate that PCML is efficient and can be implemented effectively.

INDEX TERMS Online medical diagnosis, privacy-preserving, collaborative model learning, skyline computation.

I. INTRODUCTION

In recent years, the online medical diagnosis system [1], which can provide medical diagnosis service anywhere and anytime, has attracted considerable interest. Compared with traditional treatment methods, online medical diagnosis is more flexible and convenient since it breaks the geographical restriction, and reduces the waiting time of seeing doctors [2]–[6]. To predict hidden diseases from collected medical data, many data mining techniques have been developed for e-healthcare system in recent years. For example, skyline computation [7], which returns a set of interesting points from a potentially huge data space, can be appropriately used in medical data analyzing and disease classification [6]. Specifically, with collected medical data, healthcare centers can generate diagnosis models via medical data mining with skyline query, which assists them in offering online medical diagnosis services, and allows users to check their health conditions expediently.

Unfortunately, in traditional online medical system, the medical data are commonly stored distributively in different healthcare centers, and a sole healthcare center collecting only a small set of medical data cannot generate a skyline diagnosis model accurate enough [8], [9]. For example,

The associate editor coordinating the review of this manuscript and approving it for publication was Rentao Gu.

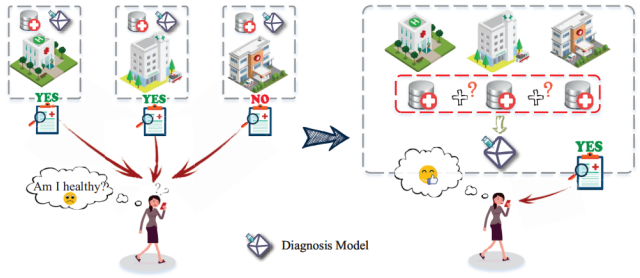


FIGURE 1. Conceptual architecture of collaborative model learning.

consider the scenario shown in Fig. 1, when a user accesses online medical diagnosis services from multiple healcare centers, due to the limitation of diagnosis model accuracy, healthcare centers may not be able to diagnose diseases accurately, which will bring bewilderment to the user. Thus, healthcare centers expect to learn a more accurate global diagnosis model collaboratively with their local medical information (i.e., local skyline diagnosis models) for offering better services.

However, owing to the data security issues, there are still many difficulties lying ahead the collaborative model learning among multiple healthcare centers [10]–[13]. In general, local skyline diagnosis models take large resources of healthcare centers to generate, and are commonly regarded as the trade secrets. Disclosure of these private information may bring an economic loss directly. As a result, healthcare centers are reluctant to contribute their local skyline diagnosis models. Therefore, it is of great importance to develop a privacy-preserving collaborative model learning scheme over multiple healthcare centers for online medical diagnosis system.

To address the above-mentioned challenges, some distributed skyline computation [14], [15] and privacy-preserving techniques [4], [16] have been proposed. Individually, distributed skyline computation techniques achieve the skyline query distributively, but these techniques mainly focus on searching the set of interesting skyline points from the distributed dataset, while not aware how to apply skyline computation in online medical diagnosis system to perform range query. Furthermore, the private information of data owners is not protected in these techniques. To address the data security issues, many privacy-preserving techniques, such as homomorphic encryption [16] and anonymity techniques [4] are proposed to achieve data security. Homomorphic encryption allows direct operations over ciphertexts, which can achieve accurate operation results on encrypted data However, most of them contain massive, complicated arithmetical operations, which brings considerable computation overhead. Anonymity techniques are extensively used in private information protection, which blurs private data into a cloaked space to protect the sensitive information, but it brings heavy communication overhead. The above-mentioned techniques can resolve the existing issues to some degree, but they are hard to be deployed in online medical diagnosis system.

In this paper, we propose a privacy-preserving collaborative model learning scheme with skyline computation, named PCML. With PCML, multiple healthcare centers can learn a global diagnosis model with their local skyline diagnosis models in the assistance of cloud, while the private local diagnosis models of healthcare centers can be well protected. Meanwhile, the learned global model is also kept confidential from the cloud. Individually, the main contributions of this paper are fourfold.

- *First*, PCML addresses the privacy and data security issues of collaborative model learning for skyline computation. With PCML, the private local skyline diagnosis models of healthcare centers are encrypted with a modified paillier cryptosystem, and are operated without decryption. Therefore, the sensitive medical information of healthcare centers can be well protected, meanwhile, the confidentiality of the final global diagnosis model is ensured.
- *Second*, PCML achieves collaborative model learning accurately. To achieve the quality of online medical diagnosis service, we construct a secure multi-party vector comparison (SMVC) algorithm based on paillier cryptosystem with secret sharing, which supports lossless collaborative model learning while protecting healthcare centers' privacy.
- *Third*, PCML accomplishes the fault-tolerant mechanism for collaborative model learning. In the real environment, the servers of healthcare centers may crash due to some irresistible factors (such as physical damage, malicious attacks). With threshold decryption technique, PCML achieves that even if a few healthcare centers are crashed, the global model can also be calculated correctly.
- *Fourth*, PCML is efficient regarding computation complexity and communication overhead. Based on manhattan distance, the skyline points can be extracted easily from datasets, which improve the efficiency of collaborative model learning. Moreover, we test PCML through PC with a real medical dataset to evaluate its effectiveness. Extensive results show that PCML is efficient and can be implemented effectively.

The remainder of this paper is organized as follows. In section II, we formalize the system model, security requirements, and identify our design goal. In section III, we review the skyline computation, bilinear pairing, and paillier cryptosystem as the preliminaries. Then, we propose our PCML in section IV, followed by the security analysis and performance evaluation in section V and section VI, respectively. We also review some related works in section VII. Finally, we draw our conclusions in section VIII.

II. MODELS AND SECURITY REQUIREMENTS

In this section, we formalize the system model and security requirements.

A. SYSTEM MODEL

In our system model, we mainly focus on how to provide privacy-preserving collaborative model learning for online medical diagnosis system. Each healthcare center is equipped with a PC, which can connect with other healthcare centers and the cloud. Specifically, the system consists of three parts: 1) trusted authority (TA); 2) healthcare centers (HCs) and 3) cloud server (CS). As shown in Fig. 2.

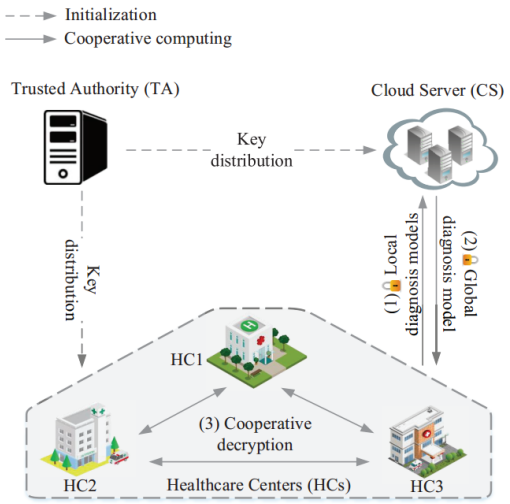


FIGURE 2. System model under considered.

- TA is a trusted authority (i.e., a government department), which bootstraps the system initialization through generating the system parameters, and distributing keys for HCs and CS.
- HCs $= \{\mathrm{HC}_1, \cdots, \mathrm{HC}_n\}$ is a set of n healthcare centers. In our system, each $\mathrm{HC}_i \in$ HCs owns a local skyline diagnosis model built upon the collected clinical datasets, and shares its encrypted local skyline diagnosis model for obtaining a more precise global diagnosis model via cooperative computing with other HCs and CS.
- CS is a cloud server which assists healthcare centers to generate global skyline diagnosis model. CS is responsible for aggregating the encrypted local diagnosis models, and generating the final global diagnosis model via cooperative computing with HCs. In our system, CS undertakes the most calculations during the collaborative model learning process.

B. THREATEN MODEL AND SECURITY REQUIREMENT

In our threaten model, we consider that CS and HCs are *honest-but-curious* [17]. Specifically, CS (1) stores the encrypted local skyline diagnosis models of HCs without tampering them; (2) honestly executes the operations of PCML, and returns the collaborative model learning result reliably; and (3) tries to retrieve the underlying plaintext of local skyline diagnosis models of HCs. In addition, each HC (1) does not send false information; (2) tries to analyze other HCs' local skyline diagnosis models during the cooperative computing process. Considering above security issues, the following security requirements should be satisfied.

- *Privacy:* Protecting the privacy of each HC's local skyline diagnosis model. Concretely, during the collaborative model learning process, every HC's local skyline diagnosis model cannot be leaked to CS and other HCs.
- *Confidentiality:* Protecting the learned global diagnosis model from the cloud. Specifically, after the collaborative model learning process, the golbal diagnosis model can only be retrieved by HCs.

III. PRELIMINARIES

In this section, we review the skyline computation and its additivity property [18], skyline diagnosis model [6], and paillier cryptosystem [19]. These will serve as the basis of our PCML.

A. SKYLINE COMPUTATION AND ADDITIVITY PROPERTY

1) SKYLINE COMPUTATION

Given a dataset S in m-dimensional space, and a point $P \in S$ can be represented as a vector $\boldsymbol{P} = \{p_1, p_2, \cdots, p_n\}$. Without loss of generality, let us assume that the p_i in any dimension i is greater or equal to zero ($p_i \geq 0$).

Definition 1 (Skyline Computation): A point $P \in S$ is said to dominate another point $Q \in S$, represented by $P \prec Q$, if it satisfies the following conditions: *1)* For every dimension i, $p_i \leq q_i$. *2)* At least there exists one dimension j such that $p_j < q_j$. The skyline set is a set of points $SKY(S) \subseteq S$ that are not dominated by any other points. The points in $SKY(S)$ are called skyline points.

2) ADDITIVITY PROPERTY

Given a dataset S and n datasets such that $S = S_1 \cup S_2 \cup \cdots \cup S_n$, the additivity of skyline computation can be represented as $SKY(S) = SKY(SKY(S_1) \cup SKY(S_2) \cup \cdots \cup SKY(S_n))$. Note that if there are same points in $SKY(S_i)$, $i = i, \cdots, n$, reserve one of them and delete the extra points. Then, the above equation implies the two following events are equivalent: *1)* Skyline computation is calculated over the union of the n datasets. *2)* Skyline computation is firstly calculated over each dataset to generate local skyline set, and then calculated over the union of local skyline sets.

Due to the additivity of skyline computation, the skyline query can be processed in a distributed method. As shown in Fig. 3, assume two datasets S_1 and S_2 stored in different data owners. The additivity of the skyline computation ensures that it is sufficient to take into account only the skyline points $SKY(S_1)$ (a, b, c, d) and $SKY(S_2)$ (e, f, g) to retrieve the global skyline points of the dataset $SKY(S) = SKY(S1 \cup S2)$ (a, e, b, g, d). This is because no other points can be part of the global skyline points since they are dominated by at least one point of $SKY(S_1)$ or $SKY(S_2)$, and the dominance relation is transitive.

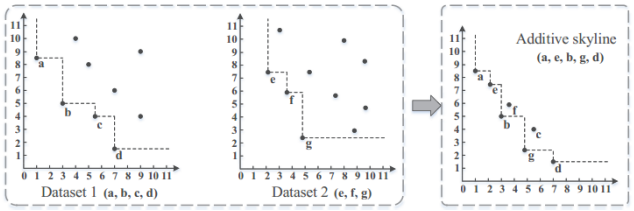


FIGURE 3. Additivity of skyline computation in two dimension.

B. SKYLINE DIAGNOSIS MODEL

From a data set S, it can be seen that the skyline points $SKY(S)$ represents the boundary of S. Through modifying the size relationship of elements in each dimension between two vector, we can define positive and negative skyline points, which present the lower boundary and upper boundary, respectively. Therefore, the skyline computation can be modified to a classifier via determining whether a tested point is within the range enclosed by the upper and lower boundaries. Then, given a medical case dataset, a disease classifier can be constructed as follows.

Assume that the medical case dataset S' in m-dimensional space with cardinality n, P and Q are two different points in S'.

Definition 2 (Positive Skyline): P is said to positive dominated Q, represented by $P \preceq Q$, if it satisfies following conditions, which are the same as $P \prec Q$: *1)* For every dimension i, $p_i \leq q_i$. *2)* At least there exists one dimension j such that $p_j < q_j$.

Definition 3 (Negative Skyline): P is said to negative dominated Q, represented by $P \succeq Q$, if it satisfies the following conditions: *1)* For every dimension i, $p_i \geq q_i$. *2)* At least there exists one dimension j such that $p_j > q_j$.

Based on above definitions, the positive points of medical dataset S', $PSKY(S')$, and the negative points of medical dataset S', $NSKY(S')$ can be calculated to construct the skyline diagnosis model. Suppose that a medical query information is presented by a point C, if C is positive dominated by at least on point in $PSKY(S)$, and is negative dominated by at least one point in $NSKY(S)$, it can be concluded that the diagnosis result of C is positive.

C. PAILLIER CRYPTOSYSTEM

Paillier cryptosystem is a widely used additive homomorphic encryption. The detail is presented as the following three functions.

1) Key Generation: $(pk, sk) \leftarrow KeyGen(\kappa)$. Choose two big primes p, q, and computes $N = pq$, $\lambda = lcm(p-1, q-1)$. Then, select a random $g \in Z^*_{N^2}$ such that $gcd(L(g^\lambda \bmod N^2), N) = 1$, where $L(x) = (x-1)/N$. The public key and the private key are $pk = (N, g)$ and $sk = \lambda$, respectively.

2) Encryption: $[\![m]\!] \leftarrow E(m, pk)$, where $E(\cdot)$ presents the encryption function of paillier. Let $m \in Z_N$ be a plaintext and $r \in Z_N$ be a random number. The ciphertext is given by $[\![m]\!] = g^m r^N \bmod N^2$.

3) Decryption: $m \leftarrow D([\![m]\!], sk)$, where $D(\cdot)$ presents the decryption function of paillier. Given a ciphertext $[\![m]\!] \in Z_{N^2}$, the corresponding plaintext can be derived as $m = (L([\![m]\!]^\lambda \bmod N^2)/L(g^\lambda \bmod N^2)) \bmod N$.

The homomorphism of paillier cryptosystem:

For any $m_1, m_2, r_1, r_2 \in Z_N$, we have additive and multiple homomorphism as follows.

$$[\![m_1]\!] \cdot [\![m_2]\!] = g^{m_1+m_2}(r_1 r_2)^N = [\![m_1 + m_2]\!],$$
$$[\![m_1]\!]^{m_2} = g^{m_1 m_2} r^{N m_2} = [\![m_1 m_2]\!].$$

IV. PROPOSED PRIVACY-PRESERVING SCHEME

In this section, we present our PCML scheme, which mainly consists of four phases: 1) *system initialization*; 2) *local diagnosis model encryption*; 3) *collaborative model learning*; and 4) *collaborative learned result reading*. The overview of PCML is described in Fig. 4. At first, TA generates system parameters, calculates the public key PK and corresponding private key SK of paillier cryptosystem, and splits the private key SK into multiple parts for HCs and CS to achieve threshold decryption. Then, HCs encrypt their local skyline diagnosis models, which will be submitted to CS later. CS executes cooperative computing with HCs to obtain the collaborative learned result, and send the result to HCs. Finally, HCs can obtain the global skyline diagnosis model via cooperative decrypting the learned result. To describe PCML more clear, we give the description of used notations in Table 1.

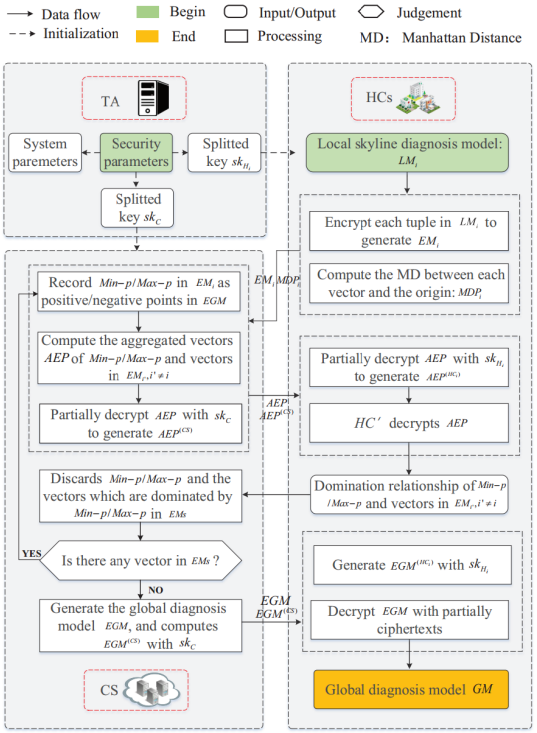


FIGURE 4. Overview of PCML.

A. SYSTEM INITIALIZATION

During the system initialization phase, TA generates the system parameters, and splits the secret key SK of paillier cryptosystem for CS and HCs.

TABLE 1. Definition of notations in PCML.

Notation	Definition
$\|m\|$	The bit length of x.
α	The security parameter.
PK, SK	The keys of paillier cryptosystem.
$\alpha_0, \cdots, \alpha_n$	$n+1$ distinct integers for splitting SK.
sk	the distributed secret keys splitted from SK.
$PSKY(\cdot)$	Positive skyline points set.
$NSKY(\cdot)$	Negative skyline points set.
LM_i	Local diagnosis model.
EM_i	Encrypted local diagnosis model.
$[\![m]\!]$	The ciphertext of x encrypted with PK.
$[\![m]\!]^{(l)}$	The partially decryption result of $[\![m]\!]$ by l.
$EPSKY(\cdot)$	The encrypted positive skyline points set.
$ENSKY(\cdot)$	The encrypted negative skyline points set.
$\pi(\cdot)$	A random permutation function.
EGM	The collaborative learned result (Encrypted global diagnosis model).
GM	Global diagnosis model.
ADS	The aggregated vector decryption set.
GDS	The global diagnosis model decryption set.

TA first chooses a security parameter α and two large safe prime numbers p, q, where $|p| = |q| = \alpha$, and computes $N = pq, \lambda = lcm(p-1, q-1)$. Then, TA chooses a generator g of order $(p-1)(q-1)/2$, such as $g = -a^{2N}$, where a is a random number in $\mathbb{Z}^*_{N^2}$ (for simplicity we denote $g = 1+N$). The public key is $PK = N$, and the corresponding private key is $SK = \lambda$.

Furthermore, TA calculates $\delta = \eta \cdot \lambda$, s.t., $\eta \cdot \lambda \equiv 1 \bmod N^2$, sets a threshold u, s.t., $u - 1 \leq n$, and defines a polynomial $q(x) = \lambda + \sum_{i=1}^{u-1} a_i x^i$, where $a_1, a_2, \cdots, a_{u-1}$ are $u-1$ random numbers from $\mathbb{Z}^*_\lambda$. Finally, let $\alpha_0, \alpha_1, \cdots, \alpha_n$ be $n+1$ distinct nonzero integers satisfying $|\varepsilon| \leq N/2$, where $\varepsilon = (max\{\alpha_0, \alpha_1, \cdots, \alpha_n\} - min\{\alpha_0, \alpha_1, \cdots, \alpha_n\})!$, TA can splits the private key SK into $n+1$ parts through computing $q(\alpha_i)$. TA publishes the system parameters $< g, PK, \eta, \alpha_1, \alpha_2, \cdots, \alpha_n, \varepsilon >$.

The private key SK is splitted to $n+1$ distributed secret keys for HCs and CS. Then, each $HC_i \in$ HCs requests TA for its distributed secret key $sk_{H_i} = q(\alpha_i)$. Similarly, CS requests TA for its distributed secret key $sk_C = q(\alpha_0)$.

B. LOCAL DIAGNOSIS MODEL ENCRYPTION

In this phase, in order to achieve the privacy of local skyline diagnosis models, every HC_i encrypts its original medical data before sending to CS.

Specifically, assume that the clinical dataset with m-dimensional, and the local skyline diagnosis model of HC_i are S_i, and LM_i, respectively. Referring to Section 3.2, LM_i is constructed by $PSKY(S_i)$ and $NSKY(S_i)$, where $PSKY(S_i) = \{\boldsymbol{P}_{i1}, \cdots, \boldsymbol{P}_{is}\}$, and $NSKY(S_i) = \{\boldsymbol{P}_{i(s+1)}, \cdots, \boldsymbol{P}_{it}\}$. In detail, LM_i can be represented with a matrix as follows.

$$LM_i = \begin{pmatrix} \boldsymbol{P}_{i1} \\ \cdots \\ \boldsymbol{P}_{is} \\ \boldsymbol{P}_{i(s+1)} \\ \cdots \\ \boldsymbol{P}_{it} \end{pmatrix} = \begin{pmatrix} p_{i11} & , \cdots , & p_{i1m} \\ \cdots & , \cdots , & \cdots \\ p_{is1} & , \cdots , & p_{ism} \\ p_{i(s+1)1} & , \cdots , & p_{i(s+1)m} \\ \cdots & , \cdots , & \cdots \\ p_{it1} & , \cdots , & p_{itm} \end{pmatrix}.$$

Then, HC_i chooses random numbers $r_{ijk} \in \mathbb{Z}_N$ (where i is the ID of HC_i, $j = 1, \cdots, t$, $k = 1, \cdots, m$), and executes the following operations to encrypt each element in LM_i with the public key $PK = N$.

$$[\![p_{ijk}]\!] = g^{p_{ijk}} \cdot r_{ijk}^N \bmod N^2. \tag{1}$$

Moreover, HC_i computes the *manhattan distance* between the origin and each vector in LM_i as follows

$$mdp_{ij} = \sum_{l=1}^{k} p_{ijk}. \tag{2}$$

After this, HC_i can obtain the encrypted local skyline diagnosis model

$$EM_i = \begin{pmatrix} \boldsymbol{EP}_{i1} \\ \cdots \\ \boldsymbol{EP}_{is} \\ \boldsymbol{EP}_{i(s+1)} \\ \cdots \\ \boldsymbol{EP}_{it} \end{pmatrix} = \begin{pmatrix} [\![p_{i11}]\!] & , \cdots , & [\![p_{i1m}]\!] \\ \cdots & , \cdots , & \cdots \\ [\![p_{is1}]\!] & , \cdots , & [\![p_{ism}]\!] \\ [\![p_{i(s+1)1}]\!] & , \cdots , & [\![p_{i(s+1)m}]\!] \\ \cdots & , \cdots , & \cdots \\ [\![p_{it1}]\!] & , \cdots , & [\![p_{itm}]\!] \end{pmatrix},$$

and the *manhattan distances* $MDP_i = (mdp_{i1}, \cdots, mdp_{it})$.

Finally, HC_i submits $< EM_i || MDP_i >$ to CS.

C. COLLABORATIVE MODEL LEARNING

After receiving total n encrypted encrypted local diagnosis model packet $< EM_i || MDP_i >$, for $i = 1, \cdots, n$. CS is responsible for extracting the positive and negative skyline points from EM_i, $i = i, \cdots, n$. Concretely, the following operations will be executed.

- ***Step-1: Encrypted Vectors Aggregation***

  Assume that each EM_i has t skyline points, CS first aggregates $MDP_1, \cdots, MDP_n$ to obtain $MDP = (mdp_{11}, \cdots, mdp_{nt})$. Then, CS marks the vectors, which is corresponding to the minimum and maximum in $(mdp_{11}, \cdots, mdp_{nt})$, as $Min - p$ and $Max - p$, respectively.

  For simplify, we take extracting the positive points from EM_i, $i = i, \cdots, n$ as an example. CS selects two random numbers $r_c, r_{c'} \in \mathbb{Z}_N$ satisfying $|r_{c'}| = \alpha/2$. Assume that $Min - p = ([\![p_{ij1}]\!], \cdots, [\![p_{ijm}]\!])$ is a vector in EM_i, and each EM_i has s positive points. Then, for each vector $\boldsymbol{EP}_{i'j'} = ([\![p_{i'j'1}]\!], \cdots, [\![p_{i'j'm}]\!])$ in $EM_{i'}$, $i' \neq i$, CS calculates

$$\begin{cases} [\![1]\!] = g \cdot r_c^N \bmod N^2 \\ [\![p'_{ijk}]\!] = [\![p_{ijk}]\!]^2 \cdot [\![1]\!] \\ [\![p'_{i'j'k}]\!] = [\![p_{i'j'k}]\!]^2 \\ acp_{i'j'k} = ([\![p'_{ijk}]\!] \cdot [\![p'_{i'j'k}]\!]^{N-1})^{\eta \cdot r_{c'}} \\ acp^{(CS)}_{i'j'k} = acp^{sk_C}_{i'j'k}, \end{cases} \tag{3}$$

  where $k = 1, \cdots, m$ and $j' = 1, \cdots, t$. After this, CS obtains the aggregated vectors $\boldsymbol{AEP}_{i'j'} = (acp_{i'j'1}, \cdots, acp_{i'j'm})$ and its partially decryption $\boldsymbol{AEP}^{(CS)}_{i'j'} = (acp^{(CS)}_{i'j'1}, \cdots, acp^{(CS)}_{i'j'm})$. Then, CS executes $\pi(\boldsymbol{AEP}_{i'j'})$ and $\pi(\boldsymbol{AEP}^{(CS)}_{i'j'})$ to make the order

of elements in $\boldsymbol{AEP}_{i'j'}$ and $\boldsymbol{AEP}_{i'j'}^{(\mathrm{CS})}$ chaotic. Moreover, CS computes $AEP = (\boldsymbol{AEP}_{11}, \cdots, \boldsymbol{AEP}_{(n-1)s})$ and $AEP^{(\mathrm{CS})} = (\boldsymbol{AEP}_{11}^{(\mathrm{CS})}, \cdots, \boldsymbol{AEP}_{(n-1)s}^{(\mathrm{CS})})$. Similarly, CS executes $\pi'(AEP)$ and $\pi'(AEP^{(\mathrm{CS})})$ to make the order of vectors in AEP and $AEP^{(\mathrm{CS})}$ chaotic.

Finally, CS sends $< AEP >$ to all HCs. Meanwhile, CS chooses a HC $\in$ HCs randomly, represented by HC$'$, and sends $< AEP^{(\mathrm{CS})} >$ to HC$'$.

- ***Step-2: Intermediate Values Calculation***

  After receiving $< AEP >$, HC$_i$ computes

$$acp_{i'j'k}^{(\mathrm{HC}_i)} = acp_{i'j'k}^{sk_{H_i}},$$

  where $i' = 1, \cdots, n-1$, $j' = 1, \cdots, t$, and $k = 1, \cdots, m$. After this, HC$_i$ obtains $AEP^{(\mathrm{HC}_i)} = (AEP_{11}^{(\mathrm{HC}_i)}, \cdots, AEP_{(n-1)s}^{(\mathrm{HC}_i)})$, where $AEP_{i'j'}^{(\mathrm{HC}_i)} = (acp_{i'j'1}^{(\mathrm{HC}_i)}, \cdots, acp_{i'j'm}^{(\mathrm{HC}_i)})$.

  Finally, each HC$_i$ (except HC$'$) sends $< AEP^{(\mathrm{HC}_i)} >$ to HC$'$.

- ***Step-3: Dominating Relationship Judgement***

  Once $< AEP >$ from CS, and $< AEP^{(\mathrm{HC}_i)} >$ from $n-1$ HCs are received, HC$'$ maps the set $\{\mathrm{CS}, \mathrm{HC}_1, \cdots, \mathrm{HC}_n\}$ to $\{0, 1, \cdots, n\}$ in order. Correspondingly, $\{acp_{i'j'}^{(\mathrm{CS})}, acp_{i'j'}^{(\mathrm{HC}_1)}, \cdots, acp_{i'j'}^{(\mathrm{HC}_n)}\}$ will be mapped to $\{acp_{i'j'}^{(0)}, acp_{i'j'}^{(1)}, \cdots, acp_{i'j'}^{(n)}\}$. After this, HC$'$ chooses arbitrary v' ($v' \geq u$) numbers from $\{0, 1, \cdots, n\}$ to construct the aggregated vector decryption set ADS, and decrypts each element in AEP through computing

$$\begin{cases} \Delta_{l,ADS}(x) = \varepsilon \cdot \prod_{l' \in ADS, l' \neq l} \dfrac{x - \alpha_{l'}}{\alpha_l - \alpha_{l'}} \\ \theta_{i'j'k} = \prod_{l \in ADS} (acp_{i'j'k}^{(l)})^{\Delta_{l,ADS}(0)} \bmod N^2 \\ \theta'_{i'j'k} = L(\theta_{i'j'k})/\varepsilon \bmod N, \end{cases} \tag{4}$$

  where $i' = 1, \cdots, n-1$, $j' = 1, \cdots, t$ and $k = 1, \cdots, m$. Then, through the value of $(\theta'_{i'j'1}, \cdots, \theta'_{i'j'm})$, HC$'$ can obtain the dominating relationship $\Theta_{i'j'}$ of $Min - p$ and $\boldsymbol{EP}_{i'j'}$ (for $k = 1, \cdots, m$, if all $|\theta'_{ijk}| > N/2$, HC$'$ can determine that $Min - p \preceq \boldsymbol{EP}_{i'j'}$, otherwise, $Min - p$ and $\boldsymbol{EP}_{i'j'}$ are irrelevant.). Moreover, HC$'$ computes $\Theta = (\Theta_{11}, \cdots, \Theta_{(n-1)t})$, which implies the dominating relationship between $Min - p$ and all vectors in $EM_{i'}$, $i' \neq i$.

  Finally, HC$'$ returns $< \Theta >$ to CS.

- ***Step-4: Skyline Points Extraction:***

  Upon receiving $< \Theta >$, CS records $Min - p$ as a encrypted positive skyline point in global diagnosis model. Meanwhile, CS discards $Min - p$ and the vectors which are positive dominated by $Min - p$ according to Θ. After this, for the remaining vectors in EM_i, $i = 1, \cdots, n$, CS repeats *step-1* to *step-4* until there is no vector in EM_i, $i = 1, \cdots, n$.

Finally, CS can obtain the encrypted positive points of global diagnosis $EPSKY(S) = (Min - p_1, \cdots, Min - p_{s'})$, where $S = S_1 \cup \cdots \cup S_n$. Furthermore, through modifying the dominating relationship from positive to negative in *step 1-4*,

Algorithm 1 SMVC: Secure Multi-Parity Vector Comparison

Input: Two encrypted vectors $\boldsymbol{EP}_{ij}$ (***Min-p*** or ***Max-p***) $= (\llbracket p_{ij1} \rrbracket, \cdots, \llbracket p_{ijm} \rrbracket)$ and $\boldsymbol{EP}_{i'j'} = (\llbracket p_{i'j'1} \rrbracket, \cdots, \llbracket p_{i'j'm} \rrbracket)$ of CS.

Output: The dominating relationship of $\boldsymbol{EP}_{ij}$ and $\boldsymbol{EP}_{i'j'}$.

```
for k = 1 to k = m do
    CS computes the aggregated element acp_ijk;
    CS computes acp_ijk^(CS) ← acp_ijk with sk_C;
    HC_i computes acp_ijk^(HC_i) ← acp_ijk with sk_Hi;
end for
CS generates AEP_ij^(CS);
HC_i generates AEP_ij^(HC_i);
HC' maps {CS, HC_1, ··· HC_n} → {0, 1, ··· , n}, and chooses ADS from {1, ··· , n};
HC' sets temp = 0;
for k = 1 to k = m do
    HC' decrypts acp_ijk to obtain θ'_ijk with ADS;
    if |θ'_ijk| > N/2 then
        HC' computes temp = temp + 1
    else
        HC' computes temp = temp − 1
    end if
end for
if temp = m then
    return EP_ij ⪯ EP_i'j';
else if temp = −m then
    return EP_ij ⪰ EP_i'j';
else
    return no relationship between EP_ij and EP_i'j';
end if
```

CS can obtain the encrypted negative points of global diagnosis $EPSKY(S) = (Max - p_1, \cdots, Min - p_{t'})$. Then, the collaborative learned result EGM (encrypted global diagnosis model) can be represented as the following matrix.

$$EGM = \begin{pmatrix} \boldsymbol{Min\text{-}p}_1 \\ \cdots \\ \boldsymbol{Min\text{-}p}_{s'} \\ \boldsymbol{Max\text{-}p}_1 \\ \cdots \\ \boldsymbol{Max\text{-}p}_{t'} \end{pmatrix} = \begin{pmatrix} eg_{11} & , \cdots , & eg_{1m} \\ \cdots & , \cdots , & \cdots \\ eg_{s'1} & , \cdots , & eg_{s'm} \\ eg_{(s'+1)1} & , \cdots , & eg_{(s'+1)m} \\ \cdots & , \cdots , & \cdots \\ eg_{t''1} & , \cdots , & eg_{t''m} \end{pmatrix}.$$

Note that each element in EGM is still the ciphertext. Then, CS computes

$$eg_{ij}^{(\mathrm{CS})} = eg_{ij}^{sk_C},$$

where $i = 1, \cdots, t''$, and $j = i, \cdots, m$, to generate $EGM^{(\mathrm{CS})}$.

Finally, CS sends the $< EGM || EGM^{(\mathrm{CS})} >$ to all HCs.

D. COLLABORATIVE LEARNED RESULT READING

After receiving the encrypted global diagnosis model packet $< EGM || EGM^{(\mathrm{CS})} >$, HC$_i$ cooperatively decrypts EGM to obtain the global diagnosis model GM.

Concretely, for every element in EGM, HC_i computes $eg_{ij}^{(\mathrm{HC}_i)} = eg_{ij}^{sk_{H_i}}$ with it distributed secret key sk_{H_i} to obtain $EGM^{(\mathrm{HC}_i)}$, and shares its $EGM^{(\mathrm{HC}_i)}$ to other HCs through a secure channel. Then, the global diagnosis model can be retrieved by HC_i with the follow steps.

- ***Step-1: Decryption Set Generation***
  HC_i maps $\{\mathrm{CS}, \mathrm{HC}_1, \cdots \mathrm{HC}_n\}$ to $\{0, 1, \cdots, n\}$ in order. Correspondingly, $\{EGM^{(\mathrm{CS})}, EGM^{(\mathrm{HC}_1)}, \cdots, EGM^{(\mathrm{HC}_n)}\}$ is mapped to $\{EGM^{(0)}, EGM^{(1)}, \cdots, EGM^{(n)}\}$. Once the numbers of HCs, which have shared their $EGM^{(\mathrm{HC})}$, is greater than $u-1$, HC_i can select v ($v \geq u$) numbers from $\{0, 1, \cdots, n\}$ to construct the global diagnosis model decryption set GDS.
- ***Step-2: Collaborative Learned Result Decryption***
  HC_i executes the follow calculations to obtain g_{ij}, which is the plaintext of each element in EGM.

$$\begin{cases} \Delta_{l,GDS}(x) = \varepsilon \cdot \prod_{l' \in GDS, l' \neq l} \dfrac{x - \alpha_{l'}}{\alpha_l - \alpha_{l'}} \\ g'_{ij} = \prod_{l \in GDS} (eg_{ij}^{(l)})^{\Delta_{l,GDS}(0)} \bmod N^2 \\ g_{ij} = L(g'_{ij})/\varepsilon \bmod N. \end{cases} \tag{5}$$

Finally, through decrypting each elements in EGM with executing the above steps, all HCs can obtain $PKY(S) = \{\boldsymbol{G}_1, \cdots, \boldsymbol{G}_{s'}\}$, and $NKY(S) = \{\boldsymbol{G}_{s'+1}, \cdots, \boldsymbol{G}_{t'}\}$ to achieve the final global skyline diagnosis model GM, which is presented as follows.

$$GM = \begin{pmatrix} \boldsymbol{G}_1 \\ \cdots \\ \boldsymbol{G}_{s'} \\ \boldsymbol{G}_{s'+1} \\ \cdots \\ \boldsymbol{G}_{t''} \end{pmatrix} = \begin{pmatrix} g_{11} & , \cdots , & g_{1m} \\ \cdots & , \cdots , & \cdots \\ g_{s'1} & , \cdots , & g_{s'm} \\ g_{(s'+1)1} & , \cdots , & g_{(s'+1)m} \\ \cdots & , \cdots , & \cdots \\ g_{t''1} & , \cdots , & g_{t''m} \end{pmatrix}.$$

Correctness of PCML: The key point of proposed PCML is to extract the positive/negative skyline points from HCs' local diagnosis models over ciphertexts, which can be verified as following.

Theorem 1: ***Min-p/Max-p*** is the positive/negative skyline point in EM_i, $i = 1, \cdots, n$.

Proof: For simplicity, we take ***Min-p*** as an example. Assume that there is another point $EP' = (\llbracket p'_1 \rrbracket, \cdots, \llbracket p'_m \rrbracket)$ that positive dominates $Min-p = (\llbracket p_{ij1} \rrbracket, \cdots, \llbracket p_{ijm} \rrbracket)$. It means $p'_1 \leq p_{ij1}, \cdots, p'_m \leq p_{ijm}$, and at least there exists one dimension k' such that $p'_{k'} < p_{ijk'}$. then, $\boldsymbol{EP'}$ will be the vector corresponding to the minimum value in $(mdp_{11}, \cdots, mdp_{nt})$, which contradicts the definition of ***Min-p***. Therefore, it can be claimed that ***Min-p*** is the positive skyline point in EM_i, $i = 1, \cdots, n$. Moreover, through modifying the dominating relationship from positive to negative, we can prove that ***Max-p*** is the negative skyline point in EM_i, $i = 1, \cdots, n$.

Theorem 2: SMVC can correctly compute the dominating relationships of two vectors over ciphertexts.

Proof: We describe the process of determining the dominating relationship of two vectors in PCML as algorithm SMVC, and verify the correctness of SMVC as follows.

First, based on the homomorphism of paillier cryptosystem, given $m \in \mathbb{Z}_N$, we have the following characteristic of $\llbracket m \rrbracket$.

$$\begin{aligned} \llbracket m \rrbracket^{N-1} &= g^{m \cdot (N-1)} \cdot r^{N \cdot (N-1)} \bmod N^2 \\ &= (1+N)^{m \cdot (N-1)} r^{N \cdot (N-1)} \bmod N^2 \\ &= (1 + (N-1)m \cdot N) \cdot r^{N \cdot (N-1)} \bmod N^2 \\ &= \llbracket -m \rrbracket. \end{aligned}$$

Then, we take $\boldsymbol{EP}_{ij} = (\llbracket p_{ij1} \rrbracket, \cdots, \llbracket p_{ijm} \rrbracket)$ and $\boldsymbol{EP}_{i'j'} = (\llbracket p_{i'j'1} \rrbracket, \cdots, \llbracket p_{i'j'm} \rrbracket)$ as inputs, and calculate equation (3) as follows.

$$\begin{aligned} acp_{ijk} &= ([[p'_{ijk}]] \cdot ([[p'_{i'j'k}]])^{N-1})^{\eta \cdot r_{c'}} \\ &= ([[2p_{ijk} + 1]] \cdot [[-2p_{i'j'k}]])^{\eta \cdot r_{c'}} \\ &= [[2(p_{ijk} - p_{i'j'k}) + 1]]^{\eta \cdot r_{c'}} \\ &= [[r_{c'} \cdot (2(p_{ijk} - p_{i'j'k}) + 1)]]^{\eta}. \end{aligned} \tag{6}$$

Moreover, in equation (4), since the number of elements in ADS is greater than the threshold u, with *shamir's secret sharing* [21], θ'_{ijk} can be calculated as follows.

$$\begin{aligned} \theta_{ijk} &= \prod_{l \in ADS} (acp_{ijk}^{(l)})^{\Delta_{l,ADS}(0)} \bmod N^2 \\ &= acp_{ijk}^{\sum_{l \in ADS} q(\alpha_l) \cdot \Delta_{l,ADS}(0)} \bmod N^2 \\ &= \llbracket r_{c'} \cdot (2(p_{ijk} - p_{i'j'k}) + 1) \rrbracket^{\varepsilon \cdot \eta \cdot \lambda} \bmod N^2 \\ &= 1 + N \cdot \varepsilon \cdot r_{c'} \cdot (2(p_{ijk} - p_{i'j'k}) + 1) \bmod N^2 \\ \theta'_{ijk} &= L(\theta_{ijk})/\varepsilon = r_{c'} \cdot (2(p_{ijk} - p_{i'j'k}) + 1) \bmod N. \end{aligned} \tag{7}$$

Note that the bit-lengths of $r_{c'}$ and N are $\alpha/2$ and 2α, respectively, and the value of medical data is not big, thus, it is obvious that SMVC satisfied the constraints $r_{c'} \cdot (2(p_{ijk} - p_{i'j'k}) + 1) < N/2$ while $p_{ijk} > p_{i'j'k}$. Therefore, if $|\theta'_{ijk}| > N/2$, it can be realized that $p_{ijk} < p_{i'j'k}$, and if $|\theta'_{ijk}| < N/2$, it can be realized that $p_{ijk} \geq p_{i'j'k}$. Finally, the dominating relationship of vectors EP_{ij} and $\boldsymbol{EP}_{i'j'}$ can be concluded.

E. SCALABILITY DISCUSSION

In PCML, we can see that the distributed secret key for each healthcare center is generated independently. That is, when a healthcare center participates the system, TA can generate a new distributed secret key for the healthcare center. Moreover, based on the fault-tolerant mechanism, even if a few healthcare centers are crashed, the collaborate computing process can also be executed normally. Therefore, it can be seen that the proposed scheme is scalable and flexible.

V. SECURITY ANALYSIS

In this section, we analyze the security of the proposed PCML. Specifically, following the security requirements discussed earlier, our analysis focuses on how to preserve the

private medical information during the collaborative model learning process.

Theorem 3: PCML achieves the privacy of HCs' local diagnosis models *LMs* and the confidentiality of global diagnosis model *GM* against *honest-but-curious* model [17]. (i.e., CS wants to obtain the underlying plaintext of *EMs* and *EGM* for stealing HCs' local diagnosis model *LMs* and the global diagnosis model *GM*, meanwhile, HCs expect to achieve each other's local skyline diagnosis models).

Proof: We illustrate that both HC's *LMs* and *GM* can be well protected during different phases of collaborative model learning.

- In the *local diagnosis model encryption* phase, all elements in $NSKY(S_i)$ and $PSKY(S_i)$ are encrypted with the public key $PK = N$, and a local skyline diagnosis model LM_i is transformed to EM_i by its owner via computing $[\![p_{ijk}]\!] = g^{p_{ijk}} \cdot r_{ijk}^{N} \bmod N^2$, where $j = 1, \cdots, t$, and $k = 1, \cdots, m$. Note that N is the public key of paillier cryptosystem with threshold decryption, and the corresponding secret key SK is splitted into $n+1$ distributed secret keys $sk_C, sk_{H_1}, \cdots, sk_{H_n}$ for CS and HCs by TA. Since a sole distributed key sk cannot retrieve the ciphertext $[\![p_{ijk}]\!]$, therefore, it is impossible for CS and a individual HC_i to decrypt the *EMs*. Moreover, since mdp_{ij} is the sum of elements in vector p_{ijk}, CS cannot retrieve the elements of p_{ijk} with mdp_{ij}. Thus, the privacy of *LMs* can be well protected in this phase.
- In the *collaborative model learning* phase, CS aggregates the encrypted vectors of *EMs* with the homomorphic characteristic of paillier cryptosystem to obtain the dominating relationship of vectors. In this process, only all HCs partially decrypt the $acp_{i'j'k}$, HC$'$ can retrieve the θ'_{ijk} to determine the dominating relationship of $Min-p/Max-p$ and $\boldsymbol{EP}_{i'j'}$. In addition, since $\theta'_{i'j'k} = L(\theta_{i'j'k}) = r_{c'} \cdot (2(p_{ijk} - p_{i'j'k}) + 1) \bmod N$, where $r_{c'}$ is a random number which is only known by CS, and CS makes the order of k and j' chaotic, thus, it is impossible for the HC$'$ to obtain the original data of *LMs*, or infer the size relationship of two compared tuples in a specific dimension. In other words, HC$'$ can only obtain the dominating relationship $Min-p/Max-p$ and $\boldsymbol{EP}_{i'j'}$, which is valueless for HC$'$ but necessary for the entire process of collaborative model learning. Thus, during this phase, the privacy of *LMs* can be verified. In addition, due to all operations in CS are over ciphertext, CS finally obtain the encrypted global diagnosis model *EGM*, which can only be decrypted by HCs. Therefore, the confidentiality of *GM* is guaranteed.
- In the *collaborative learned result reading* phase, note that all HCs decrypt the *EGM* distributedly, and each HC_i shares its partially decryption result $EGM^{(HC_i)}$. Meanwhile, *EGM* can only be retrieved while the number of HCs participating in the decryption is greater the threshold, thus, it is impossible for an individual $HC_i \in$ HCs to decrypt the final global diagnosis earlier than other HCs, which guarantees that all HCs can obtain the final global diagnosis model.

VI. PERFORMANCE EVALUATION

In this section, we first evaluate the performance of the proposed PCML in terms of the computation complexity of HCs and CS. Then, we implement PCML and deploy it with a medical machine learning dataset to evaluate its integrated performance.

A. EVALUATION ENVIRONMENT

In order to measure the integrated performance, we implement PCML on PCs with real medical datasets. Specifically, the test PCs are with 2.2 GHz six-core processor, 8 GB RAM, Windows 10, are chosen to evaluate CS and HCs, which are connected through 802.11g WLAN. Moreover, we set the bit-length of public key N is 800, and compare our PCML with the Paillier Cryptosystem-based Privacy-preserving Skyline Computation (PC-based PPSKY) [8]. The real medical datasets, which are from the UCI machine learning repository called Thyroid Disease Data (TDD) Set [22] and Heart Disease Data (HDD) Set [23], are selected to verify the effectiveness of PCML. Moreover, HDD are experimented on PCML and PC-based PPSKY to evaluate the performance of computation and communication. In detail, TDD total has 3175 tuples with 15 attributes, and HDD total has 898 tuples with 75 attributes. We choose 6 attributes which are non-boolean data in TDD, and 8 attributes that may closely related to heart diseases (such as age, blood pressure, serum cholesterol, etc.) in HDD for our simulation.

B. ACCURACY EVALUATION

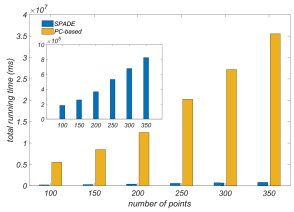
In order to verify the effectiveness of PCML, we choose 3 subsets from TDD and HDD, respectively, to simulate the clinical datasets of 3 HCs. Meanwhile, we generate the local skyline diagnosis models (LM_1, LM_2 and LM_3), and learn a global diagnosis model (GM) with the three local diagnosis models. In detail, each selected subset has 900 tuples with 6 attributes in TDD, and each selected subset has 200 tuples with 8 attributes in HDD. Moreover, we choose other 150 tuples from TDD and HDD, respectively, for testing the diagnostic accuracy of *LMs* and the *GM*. In TABLE 2, we record the number of positive/negative skyline points in LMs/GM, and the diagnostic accuracy. From the table, it can be seen that the diagnostic accuracy of global diagnosis model is much higher than local diagnosis models, therefore, with PCML, the quality of online medical diagnosis service can be greatly improved.

C. COMPUTATION COMPLEXITY

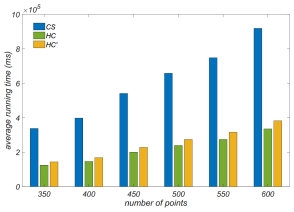
The proposed PCML can offer privacy-preserving collaborative model learning for healthcare centers, we evaluate PCML in the computation complexity of CS and HCs. For simplicity, we use HC to present a normal $HC_i \in$ HCs for distinguishing with the HC$'$ which handles some extra operations. Specifically, assume that the dimension of vectors

TABLE 2. Accuracy evaluation of TDD and HDD.

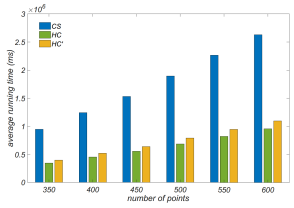
TDD	LM_1	LM_2	LM_3	GM
Positive points	482	470	457	1092
Negative points	484	473	464	1106
Accuracy(%)	73.3	68.0	69.3	87.3
HDD	LM_1	LM_2	LM_3	GM
Positive points	122	120	128	304
Negative points	123	123	129	307
Accuracy(%)	79.2	76.9	79.2	91.5



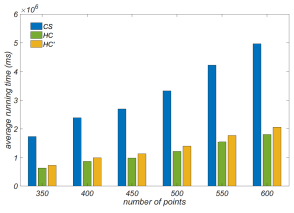
(a) Total running time of PCML vs PC-based.



(b) Average running time of PCML with dimension n=4.



(c) Average running time of PCML with dimension n=6.



(d) Average running time of PCML with dimension n=8.

FIGURE 5. Performance evaluation of computation complexity.

and the threshold are n and t, respectively. For comparing two vector, CS takes $4n$ multiplication and $6n$ exponentiation to aggregate the vectors. In addition, $2n$ multiplication and $3n$ exponentiation are taken by HC to encrypt the original data and partially decrypt the aggregated vector. Moreover, $t^2(t-1)n$ multiplications will be taken to recover the aggregated vector by HC′, or to retrieve the encrypted vector in the final learned result by HC. Denote that the multiplication operation and the multiplication operation are C_m and C_e, the total computation of CS, HC and HC′ are $4n*C_m+6n*C_e$, $n(t^3-t^2+2)*C_m+3n*C_e$, and $2n(t^3-t^2+1)*C_m+3n*C_e$, respectively.

Different from other schemes, the proposed PCML achieves privacy-preserving, multi-party model learning, high-efficiency and fault tolerate simultaneously (as shown in TABLE 3), for comparing with PCML, we select PC-based PPSKY, which is a two-part diagnosis model learing scheme relies on random permutation, 0-coding and 1-coding technique, and paillier cryptosystem. Assume that the maximum length of the binary vector which converted from original data is l, and the hash operation is C_h, then, For comparing two vectors in PC-based PPSKY, the computation of CS and HC are $4n(2s+s^2)*C_e+4ns^2*C_m$, and $24n*C_e+8n*C_m+4*C_h$, respectively.

Table 4 presents the comparison of PCML and PC-based PPSKY. We can clearly see that our proposed PCML can achieve privacy-preserving collaborative model learning with lower complexity. The factor mainly impacting the computation overhead of CS, and HCs during the collaborative model learning process is the number of points in medical dataset and the dimension of vectors, in Fig. 5(a), we plot the total running time of two-party diagnosis model learning of PCML and PC-based PPSKY with the same dimension and different number of the points. Concretely, both the datasets use 3 dimensions for testing, one of the HC uses 100 points and the number of points in another database varies from 100 to 350. From the figure, we can see that the total running time of PC-based PPSKY and PCML both increases with the number of points. However, the total running time of the PC-based PPSKY is much higher than our PCML. For testing the integrated computation overhead of PCML, we adopt that the number of HCs is 5, and in order to achieve the fault-tolerant mechanism, we set that the threshold is 4. Fig. 5(b) to (d) show the average running time of CS and HCs varying with the sum of local skyline diagnosis model points from 350 to 600, and dimension from 4 to 8, it can be clearly seen that the time of collaborative diagnosis model learning is available with real medical dataset. As a result, the above analysis of computation complexity is verified, and our proposed PCML can be applied in the real environment.

D. COMMUNICATION OVERHEAD

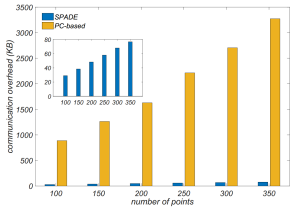
In PCML, during the *local diagnosis model encryption* phase, all HCs first submits their encrypted local model $< EM_i||MDP_i >, i = 1, \cdots, n$ to CS. Then, in the process of *collaborative model learning* phase, CS sends the encrypted aggregated vector $< AEP >$, and the partially decrypted aggregated vector $< AEP^{(\mathrm{CS})} >$ to HC and HC′, respectively. Later on, for decrypting the aggregated vector, HC sends their partially decrypted aggregated vector $< AEP^{(\mathrm{HC}_i)} >$ to HC′, and HC′ returns the dominating relationship of vectors $< \Theta >$ to CS. Finally, in the *collaborative learned result reading* phase, CS sends the encrypted global diagnosis model $< EGM||EGM^{(\mathrm{CS})} >$ to HCs, and each HC shares its $EGM^{(\mathrm{HC})}$ with other HCs. In the real environment, we record the size of these packets, and compare the total communication overhead with PC-based PPSKY in one round. Similar to computation complexity, The factor impacting the computation overhead among CS and HCs is the number of points in medical dataset and the dimension of vectors. Therefore, in Fig. 6(a), we set that one of the HC uses 100 points, and the number of points in another database varies from 100 to 350, the figure shows that with the increasing number of points, the communication overhead of PC-based PPSKY significantly increases and it is much higher than that of PCML. Furthermore, in Fig. 6(b) to (d), we adopt that the number of HCs is 5, and the threshold is 4. The figures plot the integrated communication overhead of PCML varying with the sum of local diagnosis model points from 350 to 600 and the dimension from 4 to 8, it can be

TABLE 3. Functionality comparison.

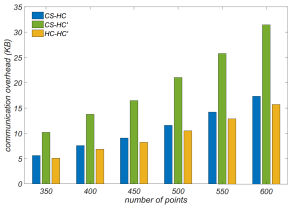
	Literature [24]	Literature [6]	Literature [25]	Literature [26]	PPSKY [8]	PCML
Multi-party learning	Yes	Yes	Yes	Yes	No	Yes
Privacy-preserving	No	No	Yes	Yes	Yes	Yes
High-efficiency	Yes	Yes	No	Yes	No	Yes
Fault tolerant	No	No	No	No	No	Yes

TABLE 4. Computation complexity of PCML vs PC-based.

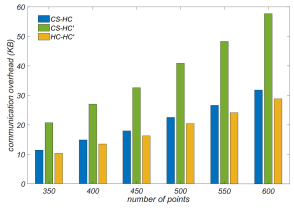
	CS	HC / HC′	Total
PCML	$4n * C_m + 6n * C_e$	$n(t^3 - t^2 + 2) * C_m + 3n * C_e$ / $2n(t^3 - t^2 + 1) * C_m + 3n * C_e$	$(8 + t^3 - t^2)n * C_m + 9n * C_e$
PC-based	$4n(2l + l^2) * C_e + 4nl^2 * C_m$	$24n * C_e + 8n * C_m + 4 * C_h$	$4n(24 + 2l + l^2) * C_e + (4l^2 + 8)n * C_m + 4 * C_h$



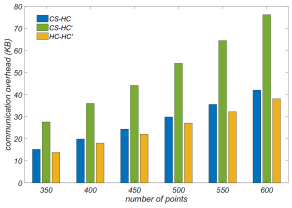
(a) Communication overhead of PCML vs PC-based.



(b) Communication overhead of PCML with dimension n=4.



(c) Communication overhead time of PCML with dimension n=6.



(d) Communication overhead of PCML with dimension n=8.

FIGURE 6. Performance evaluation of communication overhead.

clearly seen that the communication overhead is acceptable among CS and HCs in practice. In conclusion, our proposed PCML can achieve better efficiency in terms of communication overhead in CS and HCs.

VII. RELATED WORK

In this section, we briefly discuss some related works on distributed skyline computation and privacy-preserving techniques.

Skyline computation. Recently, skyline computation has received considerable attention in the database research community. Borzsonyi *et al.* [7] first presented the conception of skyline computation with algorithms called *Block Nested Loop (BNL)* and *Divide and Conquer (DC)*. Thereafter, many improved skyline computation schemes have been proposed. Papadias *et al.* [27] developed branch-and-bound skyline (BBS) based on nearest-neighbor search, which can be implemented simply in practice, and support all types of progressive processing. Zhang *et al.* [28] proposed an efficient approach to compute the skyline set using balanced partitioning, which overcomes the limitation of balancing partitioned subspaces. Lee and Huang [29] treated both dominance and incomparability as key factors in skyline computation, and proposed an algorithm named BskyTree, which outperformed traditional skyline algorithms up to two orders of magnitude. The above mentioned skyline schemes are centralized, and these schemes cannot be directly used in the distributed environment. Aiming at distributed scenarios, Vlachou *et al.* [14] addressed the efficient computation of subspace skyline queries in large-scale peer-to-peer networks, where the dataset is horizontally distributed across the peers. Specifically, based on a super peer architecture, the authors presented a threshold based algorithm, called SKYPEER, which can reduce the amount of transferred data significantly. Valkanas and Papadopoulos [15] proposed an adaptive skyline computation algorithm towards controlling the degree of parallelism and the required network traffic, in contrast to state-of-the-art methods, the proposed algorithm handles efficiently diverse preferences imposed on attributes. These schemes improved the efficiency of skyline computation, but few of them considered the data security issues.

Privacy-preserving techniques. Homomorphic encryption techniques are usual methods to achieve data operations over encrypted data without decrypting it, which can be used in privacy-preserving medical data mining. Concretely, Rahulamathavan *et al.* [5] proposed a privacy-preserving decision support system based on Gaussian kernel support vector machine (SVM) and paillier cryptosystem, which achieves that the patients' data can remain in encrypto form at all times. Liu *et al.* [2] used paillier to constructed a new cryptographic tool with additive homomorphic characteristic, and protected the sensitive medical data of users via naive Bayesian classifier over ciphertext. Kacabas and Soyata [30] presented a novel medical cloud computing approach that eliminates privacy concerns associated with the cloud provider. The proposed approach capitalizes on Fully Homomorphic Encryption (FHE), which enables computations on private health information without actually observing the underlying data. However, high time-consuming operations

are required in the most privacy-preserving schemes based on homomorphic encryption, which brings heavy computation overhead. Traditional anonymization techniques are widely used in privacy-preserving schemes such as k-anonymity and l-diversity. Belsis and Pantziou [31] presented a privacy-preserving architecture built upon the concept of k-anonymity, which allows to protect user's privacy by making an entity indistinguishable from other k similar entities. Shin *et al.* [32] proposed a novel k-member cluster seed selection algorithm based on the closeness centrality to provide consistent information loss and reduce the information distortion. Nevertheless, these anonymization techniques bring heavy communication overhead in the real environment. Furthermore, some novel privacy-preserving data mining techniques have been presented. Arefin and Morimoto [33] expanded the traditional skyline query to skyline sets queries in parallel fashion from distributed databases to protect an individual'a privacy. Liu *et al.* [8] introduced an efficient secure multi-party computation (SMC) protocol, with which the privacy-privacy skyline computation across domains can be achieved.

Different from above works, our PCML scheme aims at efficiency and privacy issues. Based on paillier cryptosystem with threshold decryption and distributed skyline computation, we develop an efficient and privacy-preserving collaborative model learning scheme for online medical system. In particular, our proposed PCML can protect healthcare centers' local diagnosis models as well as ensure the confidentiality of the final global diagnosis model, and can be easily implemented in the real environment because of its high efficiency.

VIII. CONCLUSION

In this paper, we proposed a novel privacy-preserving collaborative model learning scheme for online medical diagnosis system, called PCML. Based on paillier cryptosystem with threshold decryption and distributed skyline computation, multiple healthcare centers can securely learn a more accurate global diagnosis model with their local diagnosis models in the assistance of cloud, meanwhile, the confidentiality of the final global diagnosis model can be ensured. Specifically, before being sent to the cloud, all of the local diagnosis models are encrypted by their owner, and calculated without decryption during the collaborative model learning process. Therefore, HCs cannot obtain each other's private medical data, and CS cannot achieve any private information of HCs, as well as the final global diagnosis model. Furthermore, with threshold decryption technique, the fault-tolerant mechanism is also achieved in our scheme. Detailed security analysis shows its security strength and privacy-preserving ability, and extensive experiments were conducted to demonstrate its efficiency.

AVAILABILITY

The relevant information of the proposed scheme can be downloaded at https://www.xdzhuhui.com/demo/PCML.

REFERENCES

[1] C. A. Meier, M. C. Fitzgerald, and J. M. Smith, "EHealth: Extending, enhancing, and evolving health care," *Annu. Rev. Biomed. Eng.*, vol. 15, no. 15, pp. 359–382, 2013.
[2] X. Liu, R. Lu, J. Ma, L. Chen, and B. Qin, "Privacy-preserving patient-centric clinical decision support system on naive Bayesian classification," *IEEE J. Biomed. Health Inform.*, vol. 20, no. 2, pp. 655–668, Mar. 2016.
[3] D. V. Dimitrov, "Medical Internet of Things and big data in healthcare," *Healthcare Inform. Res.*, vol. 22, no. 3, pp. 156–163, 2016.
[4] X. Li, J. Niu, S. Kumari, J. Liao, W. Liang, and M. K. Khan, "A new authentication protocol for healthcare applications using wireless medical sensor networks with user anonymity," *Secur. Commun. Netw.*, vol. 9, no. 15, pp. 2643–2655, Oct. 2016.
[5] 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.
[6] J. Hua, H. Zhu, F. Wang, X. Liu, R. Lu, H. Li, and Y. Zhang, "CINEMA: Efficient and privacy-preserving online medical primary diagnosis with skyline query," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1450–1461, Apr. 2019.
[7] S. Borzsony, D. Kossmann, and K. Stocker, "The Skyline operator," in *Proc. 17th Int. Conf. Data Eng.*, Apr. 2001, pp. 421–430.
[8] X. Liu, R. Lu, J. Ma, L. Chen, and H. Bao, "Efficient and privacy-preserving skyline computation framework across domains," *Future Gener. Comput. Syst.*, vol. 62, pp. 161–174, Sep. 2016.
[9] C.-F. Tsai, W.-C. Lin, and S.-W. Ke, "Big data mining with parallel computing: A comparison of distributed and mapreduce methodologies," *J. Syst. Softw.*, vol. 122, pp. 83–92, Dec. 2016.
[10] S. Kim, H. Lee, and Y. D. Chung, "Privacy-preserving data cube for electronic medical records: An experimental evaluation," *Int. J. Med. Inform.*, vol. 97, pp. 33–42, Jan. 2017.
[11] 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.
[12] H. Kikuchi, T. Sato, and J. Sakuma, "Privacy-preserving hypothesis testing for the analysis of epidemiological medical data," in *Proc. IEEE 28th Int. Conf. Adv. Inf. Netw. Appl.*, May 2014, pp. 359–365.
[13] H. Zhu, F. Wang, R. Lu, F. Liu, G. Fu, and H. Li, "Efficient and privacy-preserving proximity detection schemes for social applications," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2947–2957, Aug. 2018.
[14] A. Vlachou, C. Doulkeridis, Y. Kotidis, and M. Vazirgiannis, "SKYPEER: Efficient subspace skyline computation over distributed data," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, Apr. 2007, pp. 416–425.
[15] G. Valkanas and A. N. Papadopoulos, "Efficient and adaptive distributed skyline computation," in *Proc. Int. Conf. Sci. Stat. Database Manage.* Berlin, Germany: Springer-Verlag, 2010, pp. 24–41. .
[16] M. V. Dijk, C. Gentry, S. Halevi, and V. Vaikuntanathan, "Fully homomorphic encryption over the integers," in *Proc. Annu. Int. Conf. Theory Appl. Cryptograph. Techn.*, 2010, pp. 24–43.
[17] Q. Chai and G. Gong, "Verifiable symmetric searchable encryption for semi-honest-but-curious cloud servers," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2012, pp. 917–922.
[18] K. Hose and A. Vlachou, "A survey of skyline processing in highly distributed environments," *Int. J. Very Large Data Bases*, vol. 21, no. 3, pp. 359–384, 2012.
[19] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *Proc. Int. Conf. Theory Appl. Cryptograph. Techn.*, 1999, pp. 223–238.
[20] D. Boneh and M. K. Franklin, "Identity-based encryption from the weil pairing," *SIAM J. Comput.*, vol. 32, no. 3, pp. 213–229, 2015.
[21] A. Shamir, "How to share a secret," *Commun. ACM*, vol. 22, no. 11, pp. 612–613, Nov. 1979.
[22] A. Asuncion and D. Newman. *Thyroid Disease Data Set*. Accessed: May 21, 2018. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease
[23] *Heart Disease Data Set*. Accessed: May 21, 2018. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/heart+disease
[24] H. Zhang and Q. Zhang, "Communication-efficient distributed skyline computation," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2017, pp. 437–446.
[25] Q. Wang, M. Du, X. Chen, Y. Chen, P. Zhou, X. Chen, and X. Huang, "Privacy-preserving collaborative model learning: The case of word vector training," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 12, pp. 2381–2393, Dec. 2018.

[26] Y. Zheng, R. Lu, B. Li, J. Shao, H. Yang, and K.-K. R. Choo, "Efficient privacy-preserving data merging and skyline computation over multi-source encrypted data," *Inf. Sci.*, vol. 498, pp. 91–105, Sep. 2019.
[27] D. Papadias, Y. Tao, G. Fu, and B. Seeger, "Progressive skyline computation in database systems," *ACM Trans. Database Syst.*, vol. 30, no. 1, pp. 41–82, 2005.
[28] K. Zhang, D. Yang, H. Gao, J. Li, H. Wang, and Z. Cai, "VMPSP: Efficient skyline computation using VMP-based space partitioning," in *Proc. Int. Conf. Database Syst. Adv. Appl.*, 2016, pp. 179–193.
[29] J. Lee and S.-W. Hwang, "BSkyTree: Scalable skyline computation using a balanced pivot selection," in *Proc. 13th Int. Conf. Extending Database Technol.*, 2010, pp. 195–206.
[30] O. Kocabas and T. Soyata, "Towards privacy-preserving medical cloud computing using homomorphic encryption," in *Enabling Real-time Mobile Cloud Computing Through Emerging Technologies*. Hershey, PA, USA: IGI Global, 2015, pp. 213–246.
[31] P. Belsis and G. Pantziou, "A k-anonymity privacy-preserving approach in wireless medical monitoring environments," *Pers. Ubiquitous Comput.*, vol. 18, no. 1, pp. 61–74, 2014.
[32] M. Shin, S. Yoo, K. H. Lee, and D. Lee, "Electronic medical records privacy preservation through K-anonymity clustering method," in *Proc. 6th Int. Conf. Soft Comput. Intell. Syst. 13th Int. Symp. Adv. Intell. Syst.*, 2013, pp. 1119–1124.
[33] M. S. Arefin and Y. Morimoto, "Privacy aware parallel computation of skyline sets queries from distributed databases," *Computing Informat.*, Nov./Dec. 2011, vol. 33, no. 4, pp. 186–192.

FENGWEI WANG (S'18) received the B.Sc. degree from Xidian University, Xi'an, China, in 2016, where he is currently pursuing the Ph.D. degree with the School of Cyber Engineering.

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

HUI ZHU (M'13–SM'18) received the B.Sc. degree from Xidian University, in 2003, the M.Sc. degree from Wuhan University, in 2005, and the Ph.D. degree from Xidian University, in 2009.

In 2013, he was with the School of Electrical and Electronics Engineering, Nanyang Technological University, as a Research Fellow. Since 2016, he has been a Professor at the School of Cyber Engineering, Xidian University, China. His research interests include applied cryptography, cloud security, and big data security and privacy.

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

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

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

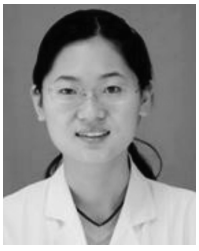
He has been an Assistant Professor with the Faculty of Computer Science, University of New Brunswick (UNB), Canada, since August 2016. Before that, he was an Assistant Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore, from April 2013 to August 2016. He was a Postdoctoral Fellow with the University of Waterloo, from May 2012 to April 2013. His research interests include applied cryptography, privacy enhancing technologies, and the IoT-big data security and privacy. He received the most prestigious Governor General's Gold Medal and the 8th IEEE Communications Society (ComSoc) Asia Pacific (AP) Outstanding Young Researcher Award, in 2013. He is currently a Senior Member of the IEEE Communications Society. He currently serves as the Secretary of the IEEE ComSocCIS-TC.

JIAFENG HUA (S'18) received the B.Sc. degree from the North University of China, Taiyuan, in 2012. He is currently pursuing the Ph.D. degree with the School of Cyber Engineering, Xidian University, Xi'an, China.

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

HUI LI (M'10) received the B.Sc. degree from Fudan University, in 1990, the M.Sc. and Ph.D. degrees from Xidian University, in 1993 and 1998, respectively.

Since 2005, he has been a Professor with the School of Telecommunication Engineering, Xidian University, China. His research interests include cryptography, wireless network security, information theory, and network coding. He served as the TPC Co-Chair for ISPEC 2009 and IAS 2009, the General Co-Chair for E-Forensic 2010, ProvSec 2011, and ISC 2011, and the Honorary Chair for NSS 2014 and ASIACCS 2016.



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

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