Privacy-preserving disease detection techniques in e-healthcare systems

Wang, Guoming

2020

Wang, G. (2020). Privacy-preserving disease detection techniques in e-healthcare systems.
Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/138085>

<https://doi.org/10.32657/10356/138085>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0
International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 15:25:39 SGT

Privacy-Preserving Disease Detection Techniques in e-Healthcare Systems

Wang Guoming

School of Electrical and Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

2020

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

22-Aug-19

.....
Date

WANG Guoming
[Student's Signature Here]

.....
[Student's Name Here]

WANG Guoming

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

22 - Aug - 2019

[Date Here]

.....
Date

[Supervisor's Signature Here]

.....
[Supervisor's Name Here]

Guan Yong Liang

Authorship Attribution Statement

This thesis contains material from 3 paper(s) published in the following peer-reviewed journal(s) where I was the first author.

Chapter 3 is published as Wang G, Lu R, Huang C, Guan YL. An efficient and privacy-Preserving pre-clinical guide scheme for mobile eHealthcare. Journal of Information Security and Applications. 2019 Jun 1;46:271-80.

The contributions of the co-authors are as follows:

- I and Prof. Lu Rongxing provided the initial project direction.
- I discussed the idea with Huang Chen, Huang Chen gave me some suggestions.
- I prepared the manuscript drafts. The manuscripts was revised by Prof. Lu Rongxing and Prof. Guan Yong Liang.
- I designed the performance evaluation, coded the programs, and analyzed the result.

Chapter 4 is published as Wang G, Lu R, Guan YL. Enabling efficient and privacy-preserving health query over outsourced cloud. IEEE Access. 2018 Nov 9;6:70831-42.

The contributions of the co-authors are as follows:

- I and Prof. Lu Rongxing provided the initial project direction.
- I prepared the manuscript drafts. The manuscripts was revised by Prof. Lu Rongxing and Prof. Guan Yong Liang.
- I designed the performance evaluation, coded the programs, and analyzed the result.

Chapter 5 is published as Wang G, Lu R, Guan YL. Achieve Privacy

eHealthcare System. IEEE Access. 2019 Jan 9;7:33565-76.

The contributions of the co-authors are as follows:

- I and Prof. Lu Rongxing provided the initial project direction.
- I prepared the manuscript drafts. The manuscripts was revised by Prof. Lu Rongxing and Prof. Guan Yong Liang.
- I designed the performance evaluation, coded the programs, and analyzed the result.

22-Aug-19

.....
Date

WANG Guoming
[Student's Signature Here]

.....
[Student's Name Here]

WANG Guoming

Acknowledgments

This thesis would not have been possible without the help and support of many people throughout my life. I express my deep appreciation to these people, regardless of whether they are mentioned here.

First of all, I would like to express my sincere gratitude to my supervisor, Prof. Guan Yong Liang, for the opportunity provided to me to pursue my PhD at Nanyang Technological University (NTU) and for his patient guidance, encouragement, and insightful suggestions throughout my PhD study and research life.

Secondly, I would like to express the depth of my gratitude to my co-supervisor, Prof. Lu Rongxing at University of New Brunswick (UNB), Canada - for his invaluable guidance and long lasting encouragement. The PhD studies with him stimulated my interests in exploring the unknowns. Moreover, Prof. Lu not only teaches me how to conduct the rigorous scholarship, but also advises me to focus on the greatest thing in work and in life. I am so fortunate to meet this outstanding supervisor. His hard working attitude and rigorous attitude of scholarship have significantly influenced on my future career.

In addition, I thank the staff members and students in my research laboratory INFINITUS, who have helped me both in study and life in Singapore, including Dr. Wang Wei, Dr. Luo Sheng, Dr. Zhang Hen, Dr. Hu Hao, Dr. Li Xiang, Dr. Huang Jiajia, Dr. Liu Aifei, Dr. Yang Rong, Dr. Liu Ximeng, Dr. Chen Le, Dr. Li Chen, Dr. Che Yueling, Dr. Fan

Rui, Dr. Li Ruoyu, Mr. Wang Yang, Mr. Liu Yunxiang, Mr. Li Beibei, Mr. Pan Zihan, Mr. Shao Hongxin, Ms. Sun Meng, and Ms. Gao Yumeng. If anyone has been left out, please accept my apologies.

Most of all, I am grateful to thank Dr. Qi Huibo; Discussions with her really inspire me to rethink my problems from research and life perspectives. I would not have been able to complete my PhD degree without her support.

Abstract

With the increasing popularity of pervasive devices such as smartphones, Body Sensor Network(BSN), Internet-of-Things devices and cloud computing, mobile e-healthcare has become a research trend in recent years. Disease detection using big data analytics techniques is a popular e-healthcare research focus. However, big-data disease detection still faces many challenges on privacy of users' sensitive personal information, confidentiality of health service provider's diagnosis model, accuracy of diagnosis, efficiency of query, etc. In this thesis, we aim to improve the security and privacy performance of the e-healthcare systems, while achieving efficient disease detection.

Firstly, we propose an efficient privacy-preserving pre-clinical guidance service scheme (PGuide) to provide on-the-go medical guidance service while preserving user privacy. Using the PGuide scheme, users can personally conduct privacy-preserving pre-clinical diagnosis based on their health profiles and obtain recommendation from trusted sources (e.g. hospitals and medical service providers) based on the diagnosis. In addition, the information transmitted to the hospitals and other medical service providers to calculate the disease risk use a disease prediction model in a privacy-preserving way.

Secondly, we propose an efficient privacy-preserving health query scheme over outsourced cloud (HeOC). In this scheme, the user shortlists the possible disease on his own using self-collected health sensor data first. Then the user can query the disease severity

accurately from the filtered result in the cloud. To reduce the query latency, we propose a novel sensor anomaly detection technique (SADS) for detecting high-risk disease. In the SADS technique, the health service provider outsources an encrypted health tree to the cloud. Authenticated users send encrypted physiological data to the cloud to detect high-risk disease without disclosing his/her sensitive personal health data. Then, with the oblivious pseudorandom function protocol (OPRF), the user retrieves the diagnosis result accurately.

Thirdly, we propose an efficient and privacy-preserving priority classification scheme (PPC), for classifying encrypted patient data at the Wireless Body Area Network gateway (WBAN-gateway) in a remote e-healthcare system. Specifically, to reduce the system latency, we design a non-interactive privacy-preserving priority classification algorithm, which enables the WBAN-gateway to perform privacy-preserving priority classification on the received users medical packets by itself and relay these packets to the healthcare center according to their criticality, i.e. more critical cases get reported first.

Contents

Acknowledgments	i
Abstract	iii
Contents	ix
List of abbreviations	x
List of figures	xiii
List of tables	xiv
1 Introduction	1
1.1 E-Healthcare Systems	1
1.1.1 Background	1
1.1.2 Applications of e-Healthcare Disease Detection Systems	2
1.1.3 Challenges for e-Healthcare system	3
1.1.4 Security Requirements	4
1.2 Research Motivations and Objectives	5
1.2.1 Motivations	5

CONTENTS

1.2.2	Objectives	6
1.3	Research Contributions	8
1.4	Outline of the Thesis	9
2	Literature Review and Fundamental Concepts	11
2.1	Literature Review	11
2.1.1	Pseudonymization based schemes	12
2.1.2	Access control based schemes	13
2.1.3	Data encryption based schemes	13
2.1.4	Differential privacy based schemes	14
2.1.5	Disease Risk Detection	15
2.2	Fundamental Concepts	15
2.2.1	Bilinear Pairings	15
2.2.2	Bilinear Pairing with Composite Order	16
2.2.3	BGN Homomorphic Cryptosystem	17
2.2.4	Max heap	18
2.2.5	Pseudorandom Functions	19
2.2.6	IND-CCA2 Secure Public-Key Encryption technique	20
2.2.7	Disease Risk Model	22
2.2.8	Determination of Disease Risk Threshold	23
3	An Efficient and Privacy-Preserving Pre-Clinical Guide Scheme for Mobile e-Healthcare	24
3.1	Introduction	24
3.2	Models and Design Goal	25

3.2.1	System Model	25
3.2.2	Security Model	27
3.2.3	Design Goal	28
3.3	Proposed PGuide Scheme	29
3.3.1	System Setting	29
3.3.2	Privacy-preserving Pre-clinical Guidance	30
3.4	Security Analysis	38
3.5	Performance Evaluation	42
3.5.1	Pre-disease Diagnosis	42
3.5.2	Recommendation of Hospitals	45
3.6	Conclusions	48
4	Efficient and Privacy-Preserving Health Query over Outsourced Cloud	49
4.1	Introduction	49
4.2	Models and Design Goal	53
4.2.1	System Model	53
4.2.2	Security Requirement	54
4.2.3	Design Goal	56
4.3	Proposed HeOC Scheme	56
4.3.1	Overview	57
4.3.2	Proposed D-Merkle Tree	59
4.3.3	System Setup	61
4.3.4	Sensor Anomaly Detection Technique	64
4.3.5	Privacy-preserving Disease Level Query	66
4.3.6	Correctness Analysis	68

4.4	Security Analysis	70
4.5	Performance Evaluation	71
4.5.1	Implementation and Experimental Settings	72
4.5.2	Hospital's computational cost	72
4.5.3	Computational cost at user side	73
4.5.4	Computational cost at cloud side	75
4.5.5	Communication Cost	76
4.6	Conclusions	77
5	Privacy-Preserving Priority Classification on Patient Health Data in Remote e-Healthcare System	78
5.1	Introduction	79
5.2	Models and Design Goal	82
5.2.1	System Model	82
5.2.2	Security Requirement	84
5.2.3	Design Goal	86
5.3	Proposed Privacy-Preserving Priority Classification Scheme	87
5.3.1	Overview	87
5.3.2	System Setup	89
5.3.3	User Registration	92
5.3.4	User Data Collection	93
5.3.5	WBAN-Gateway Packet Classification and Relay	94
5.3.6	Correctness Analysis	95
5.4	Security Analysis	98
5.5	Performance Evaluation	100

5.5.1	Implementation and Experimental Settings	101
5.5.2	Pailliar-based privacy-preserving priority classification scheme . . .	101
5.5.3	Healthcare center's computational cost	102
5.5.4	Computational cost at user side	103
5.5.5	Computational cost at WBAN-gateway side	105
5.5.6	Communication Cost	107
5.6	Conclusions	108
6	Conclusions and Future Work	109
6.1	Conclusions	109
6.2	Future Work	111
	List of Publications	114
	Bibliography	115

List of Abbreviations

AES	Advanced Encryption Standard
BSN	Body Sensor Network
CDH	Computational Diffie-Hellman
DDH	Decisional Diffie-Hellman
FHE	Fully Homomorphic Encryption
HeOC	Health Query over Outsourced Cloud
HIPAA	Health Insurance Portability and Accountability Act of 1996
HIS	Hospital Information System
IP	Intellectual Property
IND-CCA2	Adaptive Chosen Ciphertext Attack
JPBC	Java Pairing-Based Cryptography Library
OPRF	Oblivious Pseudorandom Function Protocol
OR	Odds Ratio
PGuide	Pre-Clinical Guide Scheme
PHE	Partial Homomorphic Encryption
PHI	Protected Health Information
PKE	Public-key Encryption Technique
PPC	Privacy-preserving priority classification

PPCP	Privacy-preserving Comparison Protocol
PRF	Pseudorandom Function Protocol
SADS	Sensor Anomaly Detection Technique
SHA	Secure Hash Algorithm
SP	Service Provider
TA	Trust Authority
WBAN	Wireless Body Area Network

List of Figures

2.1	A sample of max heap	19
3.1	System model under consideration	26
3.2	PPCP protocol	32
3.3	Prototype: Android app for Pre-disease Diagnosis	43
3.4	Latency in the pre-disease diagnosis process for our proposed pGuide and paillier-based scheme [1]	45
3.5	Recommendation of hospitals Android app	45
3.6	Comparative Summary: Key generation time, key sizes for user-side and communication cost with varying number of hospitals	47
4.1	Wearable health monitor system	50
4.2	System model of outsourced health query	54
4.3	Data discretization to turn numerical data into categories.	57
4.4	Query flow in HeOC	58
4.5	An example of Disease Merkle Tree with physiological data $x_1, \dots, x_{10}$, which are collected from sensors $s_1, \dots, s_{10}$	60
4.6	Data structures of nodes for D-Merkle tree outsourced to cloud	61

4.7	Hospital latency for building the <i>hdmt</i>	73
4.8	User latency for downloading diseases.	74
4.9	User verification latency and related <i>hdmt</i> tree level.	75
4.10	Cloud latency for the user to query one disease	76
5.1	Wearable health monitor system	80
5.2	System model of medical packet classification and relay in remote e-healthcare system under consideration	83
5.3	Two samples of the priority calculation with <i>sum</i> and the thresholds $(t_1, t_2, \dots, t_l)$	88
5.4	Data flow of the PPC scheme.	90
5.5	Paillier-based priority classification scheme similar to the work [2]	102
5.6	Computational cost of healthcare center for PPC and Paillier-based scheme [2]	104
5.7	Latency for user data collection for PPC and Paillier-based scheme [2]	105
5.8	Latency for priority classification for PPC and Paillier-based scheme [2]	106

List of Tables

3.1	Disease coefficients and the question set.	30
3.2	Parameters setting for PPCP	32
3.3	Experiment setup	43
4.1	Notations frequently used in HeOC	59
4.2	Experimental setting	72
4.3	Parameter setting	72
5.1	Notations used in PPC	91
5.2	Experimental setting	101

Chapter 1

Introduction

1.1 E-Healthcare Systems

1.1.1 Background

One of the most pressing issues facing hospital administration is the shortage of medical doctors, particularly with the increasing number of patients and in an ageing population. This has resulted in an increased waiting time for patients in many countries. For example, a patient has to wait between 4 and 24 hours to be seen in a hospital in Canada [3], and significantly longer in China [4,5]. Patients are unlikely to be familiar with medical departments in hospitals and most are certainly not familiar with the symptoms associated with the different diseases. For example, there are approximately 60 different types of doctors and specialists [6] in a typical hospital, it would be a frustrating and time-wasting exercise if a patient consults a doctor not trained or specialized in the particular disease. Some major hospitals have established manned enquiry counters to guide patients to go to choose the right department or specialist, but this is not a viable solution due to a number of reasons.

For example, a patient is unlikely or willing to share sensitive health-related issues, such as HIV and mental illness, publicly over the manned enquiry counters. Therefore, we need an effective solution to help ease the stretch on the limited medical resources in hospitals.

Due to the increasing digitization of our society and popularity of pervasive devices (e.g. smartphones), where most of our data (including healthcare-related data) are available electronically, using big data analytics to solve healthcare-related issues (e.g. disease risk assessment) has been the subject of recent research focus. Moreover, with the pervasiveness of smartphones and the wireless body area network (WBAN), the remote e-Healthcare system has received considerable attention.

However, the security and privacy issues have to be considered while developing the e-Healthcare systems. For example, how can we be assured that, without our consent, our sensitive medical information are not made available and exploited by third-party companies, such as insurance companies and future employers? In addition to the privacy requirements of medical users, information leakage is a major concern for service providers, as medical providers in countries such as U.S. are subject to exacting regulatory regime (e.g. HIPAA [7]).

1.1.2 Applications of e-Healthcare Disease Detection Systems

A disease detection e-Healthcare system is comprised of entities like users, gateways, cloud, service provider and healthcare center. Considering different system models, there are many applications of the disease detection e-Healthcare systems.

- Pre-clinical diagnosis: Before users go to see the doctors, they can personally conduct pre-clinical diagnosis based on their health profiles [8] and obtain recommendation from trusted sources (e.g. hospitals and medical service providers) based on the

diagnosis. It is also very helpful to make use of this pre-clinical diagnosis system to fulfill the hierarchical medical system [9].

- Elder-assistance: When an elderly person is equipped with body sensors and a smart-phone, he/she can get remote monitoring e-Healthcare service from the healthcare center. Smart devices like smart phones and smart watches can be used to collect these sensor data, conduct diagnosis [10–12]. The elder-assistance system can detect the elderly person's health condition and alert if accidents happen.

1.1.3 Challenges for e-Healthcare system

We outline the challenges that a privacy-preserving remote e-Healthcare system may face:

- Challenges on security and privacy. All the users' physiological data, personal information and the healthcare center's disease models need to be encrypted. It should not be possible for an attacker to recover the sensitive plaintext by observing the ciphertext. Moreover, it is reasonable to assume that in some scenarios, the attacker knows some users' information or some disease models. Even in this context, it should not be possible for an attacker to recover other plaintext of the corresponding encrypted data.
- Challenges on accuracy. To meet the security requirements of the remote e-Healthcare system, some privacy-preserving schemes based on the encrypted data need to standardize the users' personal information and healthcare center's disease model first. The standardization techniques may compromise the computational accuracy. What's more, some randomization techniques like the differential privacy [13] add random

values to the computational results, which may cause medical disaster in some scenarios [14]. Therefore, the privacy-preserving remote e-Healthcare system should be accurate for medical analysis.

- **Challenges on computational efficiency.** Most of the privacy-preserving schemes based on the encrypted data have high computational overhead. Recent studies propose privacy-preserving schemes with multiple parties [15], which introduce high communication cost. On the other hand, non-interactive privacy-preserving schemes are always associated with time-consuming techniques, such as fully homomorphic encryption [16]. Thus, the privacy-preserving remote e-Healthcare system needs to address the computational efficiency issues.

1.1.4 Security Requirements

In order to protect the disease detection e-Healthcare system, the security mechanisms employed in the e-Healthcare system should satisfy the following requirements:

- *Privacy.* In the health query scheme over outsourced cloud (HeOC), each user conducts health queries with sensor collected physiological and personal information like age, gender, medical history. These sensitive personal data should be prevented from being leaked to the cloud. Specifically, it should be not possible for the cloud to know the user's physiological data by observing the ciphertext of the query data, which means that the HeOC scheme is secure under ciphertext-only attack. Moreover, if the attackers know the plaintexts of some encrypted user query data, it should not be possible for the attackers to reveal the user's physiological data corresponding to the other ciphertexts. In other words, the healthcare system is secure under known-plaintext attack.

- *Confidentiality.* The diagnosis models should be prevented from being disclosed to the cloud, since these models are the intellectual properties belong to the hospital. Same as above, it should not possible for the cloud to recover the diagnosis model with the encrypted medical data outsourced from the hospital, which means that the system should be secure under ciphertext-only attack. What's more, even if the attacker knows the plaintext of one encrypted disease diagnosis model, it should be not possible for the attacker to obtain the plaintext of the encrypted diagnosis model. The system should be secure under known-plaintext attack.
- *Authentication.* Only authenticated users are allowed to send health query to the cloud and get the diagnosis result. If an illegal user forges a health query, this malicious operation should be detected immediately.

1.2 Research Motivations and Objectives

1.2.1 Motivations

The motivations of this research are highlighted as below:

- *User's privacy and system security in two-party online health query system.* An era of open information in healthcare is now under way. Healthcare experts, service providers and other practitioners have stepped up to the plate and analyze big data to obtain insights. Although these efforts are still in their early stages, they could collectively help the industry address problems relevant to the variability in healthcare quality and escalating healthcare spending. For instance, researchers can mine the data to see what treatments are most effective for particular conditions, gain important information that can help patients and reduce costs. In two-party online health

query system, each user offers his personal information to the healthcare center. To protect the user's privacy, the personal information would be encrypted first. The healthcare center needs to conduct the disease diagnosis with the encrypted user's personal information.

- *User's privacy and healthcare center's data confidentiality in cloud-based online health query system.* With the pervasiveness of smartphones, wireless body area networks and cloud computing, cloud-based remote e-Healthcare system has received considerable attentions and become more popular. A variety of BSN applications are proposed in these years including early detection, elderly assistance, physical activity monitoring and so on [17–19]. In the cloud-based online health query system, healthcare center outsources the disease diagnosis tasks to the cloud service provider. Each user offers his sensor collected information to the cloud service provider. Because the cloud service provider is semi-trust, the user's privacy and healthcare center's data confidentiality should be preserved.
- *Computational efficiency and user's privacy in information relay system.* In some information relay systems, the relay channel is semi-trust. Thus, the user's medical packets are encrypted before being relayed to healthcare center. Considering the emergency of the some medical systems, such as elderly-assistance system, the relay channel would report the medical packets with high severity firstly. Therefore, the computational efficiency issue and the user's privacy issue need to be addressed.

1.2.2 Objectives

As discussed above, since the security and privacy issues are considered in the e-Healthcare systems, the study of the privacy-preserving e-Healthcare system becomes promising. A

variety of privacy-preserving disease detection protocols and data analysis remote e-Healthcare systems can be developed. Among them, more challenging security and privacy issues are also reared up according to system models. For example, in some e-Healthcare system, the semi-honest service provider (cloud) is involved to make the security and privacy issues more complicated.

In this thesis, we will study different privacy-preserving disease detection e-Healthcare systems for different system models:

- In some two-party system models, which only consist of users and healthcare center, each user sends health query to the healthcare center. In this scenario, we focus on designing a two-party online query system between the users and the healthcare center, while preserving the users' personal information privacy and the system security.
- In some cloud-based multi-party system models, which consist of users, healthcare center and the service provider (cloud), healthcare center outsources the disease detection service to the service provider (cloud), and the users conduct health query from the service provider (cloud). In this scenario, we focus on designing a cloud-based multi-party online disease diagnosis system, while preserving the user's personal information privacy, healthcare center's data confidentiality and efficiency of the systems.
- In the healthcare information relay system models, which consist of users, healthcare center and gateways, the users conduct the health query from the healthcare center. In the query procedure, the query information packets are relayed by the gateways. We let the gateways conduct some preprocessing to make the e-Healthcare more efficient. In this context, we focus on designing a medical packets relay protocol for

emergency medical system, while preserving the user's personal information privacy, healthcare center's data confidentiality and efficiency of the systems.

1.3 Research Contributions

The research in this thesis focuses on developing a suite of protocols to deal with the challenging security and privacy-preserving issues in disease detection e-Healthcare systems. Specifically, the main contributions lie in the following aspects:

- We propose an efficient pre-clinical guidance service scheme (PGuide) [20] designed to provide on-the-go medical guidance service while preserving user privacy. With PGuide, users can personally conduct privacy-preserving pre-clinical diagnosis based on their health profiles [8], and obtain recommendation from trusted sources (e.g. hospitals and medical service providers) based on the diagnosis. The information transmitted to the hospitals and other medical service providers to calculate the health risk use a disease prediction model in a privacy-preserving way. We prove that our proposed PGuide scheme achieves the privacy-preservation for both the individual user and the medical service provider. To demonstrate the practicality of our scheme, we develop an Android app and a Java service application. Based on the findings from our evaluations, we show that our proposed PGuide scheme is efficient, in terms of computational cost and communication overhead.
- We propose an efficient and privacy-preserving health query scheme over the out-sourced cloud, named HeOC [21], which allows authenticated users to query on the encrypted health data without disclosing his/her sensitive physiological data to the cloud. Specifically, we design a novel sensor-based anomaly detection technique,

to detect the health anomaly with high risk for further analysis. Particularly, in this scheme, the user detects the anomaly disease with the sensor collected data first to filter out the suspicious disease with its related physiological data. Then through a variant of the OPRF protocol, the user can query the severity level of his health problem using the filtered result. Experiments demonstrate that our HeOC scheme is efficient. Secondly, we propose a novel sensor-based anomaly detection technique named SADS for detecting high-risk disease. In the SADS technique, the health service provider outsources an encrypted health tree to the cloud. Authenticated users send encrypted physiological data to the cloud to detect high-risk disease without disclosing his/her sensitive personal health data.

- We propose an efficient and privacy-preserving priority classification technique (PPC) for patient health data in remote e-Healthcare system [22], which allows authenticated users to periodically send medical packets to the healthcare center through WBAN-gateways. The WBAN-gateways relay these aggregated medical packets in a non-interactive privacy-preserving way based on the packets' priorities (criticalities). Particularly, The WBAN-gateways derive the priorities of the medical packets and relay the packets in a priority heap.

1.4 Outline of the Thesis

The organization of the remainder of the thesis is as follows. Chapter 2 recalls literature review and some fundamental concepts. Chapter 3 presents an efficient and privacy-preserving pre-clinical guidance scheme (PGuide). Chapter 4 presents an efficient and privacy-preserving health query over outsourced cloud (HeOC), which preserves user's

1.4. OUTLINE OF THE THESIS

personal privacy and achieves the confidentiality of the health service provider's diagnosis model. Chapter 5 presents an efficient privacy-preserving priority classification (PPC) scheme on patient healthcare data in remote e-Healthcare system. The proposed PPC scheme achieves the priority classification and packets relay tasks, while preserving the privacy of the users and the confidentiality of the healthcare center's disease models. Finally, conclusions and future research work are described in Chapter 6.

Chapter 2

Literature Review and Fundamental Concepts

2.1 Literature Review

There are many techniques involved in achieving the privacy-preserving e-Healthcare system, such as pseudonymization, data encryption, access control, privacy-preserving data outsourcing. We will review four kinds of privacy-preserving techniques used for privacy-preserving e-Healthcare system in this chapter: 1) pseudonymization; 2) access control; 3) data encryption; 4) differential privacy. Moreover, we will review some works on disease detection. Next, we will review the works applying these techniques to achieve privacy-preserving e-Healthcare system.

2.1.1 Pseudonymization based schemes

Pseudonymization is the earliest technique used for privacy-preserving e-Healthcare system. Many protocols around pseudonymization [23–27] have been studied in recent years. The key idea is to remove all the information that can identify the users. The real identity is replaced by the pseudonym before data sharing or data publishing. The attacker can not link the pseudonym to the patients. Proposals [28, 29] categorize the patients' data into two sets: user-relevant data and personal, pseudonymized data. These schemes not only deny any link between the pseudonym and the real users, but also securely store tabled entities of these identities. The pseudonymization solved the privacy concern in the early stage of the e-Healthcare system, when the cloud computing based architecture is not pervasive. However, when data information are aggregated from different data sources, the pseudonymization itself is a weak protection technique for user's privacy.

K-anonymity is a notion defining a database. Firstly we introduce the notation quasi-identifiers, i.e., attributes that can be exploited for linking the records to the users in the database. K-anonymity means that the quasi-identifiers of every record in a database are related to no fewer than k users. Many applications and techniques [30–35] are proposed to achieve k -anonymity. A paper [31] proposes an application of k -anonymity to preserve privacy in wireless sensor network medical environments. Another paper [30] presents a k -member cluster seed selection algorithm (KMCSSA) and applies it in k -member clustering to achieve k -anonymity. Recently, a proposal [35] presented a scalable k -anonymization approach using MapReduce for very large medical databases. K-anonymity is necessary in the case of data sharing to preserve privacy. However, if attackers reveal the information of one user related to some quasi-identifiers, all the users from the same quasi-identifiers will be identified.

2.1.2 Access control based schemes

The access control policies are techniques used for privacy preservation. Most of time, hybrid access control policies are widely used to propose privacy-preserving access control mechanisms [36–41]. It is common to use a combination of access control and pseudonymization in one privacy-preserving scheme, which stores the users' data in an anonymized manner, and shares the anonymized data according to some access control policies. A patient monitoring scheme [42] was proposed to give patients control over who can access their protected health information (PHI). The patients assign various categories of access to their PHI after signing the contract with the healthcare center regarding use of their PHI. A cloud-centered privacy-aware role based access control (CPRBAC) mechanism [43] was proposed to improve the traditional RBAC. It comprises of not only the context-based access control, information sharing among different cloud servers, and authorization delegation from the traditional RBAC, but also four new conditions: purpose, obligations, conditions, organizations to define complex access control policies.

2.1.3 Data encryption based schemes

Data encryptions are widely used in the privacy-preserving e-Healthcare schemes. Some pseudonymization schemes encrypt the real identities of the user record as the pseudonyms. Encryption are also widely used in access control, such as identity-based encryption in the identity-based access control or role-based access control (RBAC) scheme [44]. Moreover, a number of privacy-preserving eHealthcare system are proposed on encrypted patient data [45–52].

There are two kinds of homomorphic encryption, which include partial homomorphic encryption (PHE) [53, 54] that allows addition of encrypted data, and fully homomorphic

encryption (FHE) [16, 55] that allows both addition and multiplication on encrypted data. ElGammal [56] and Paillier [57] are most widely used to implement homomorphic encryption systems [1, 58–67]. Specifically, the fully homomorphic encryption achieves the health query on encrypted data. However, its implementation involves of high computation overhead. A recent work [68] proposed a novel parallelization to speed up the fully homomorphic based scheme. How to make scheme efficient is a big challenge while adopting homomorphic encryption, especially in large health databases.

Recently, many papers [60, 69, 70] have studied the privacy-preserving health monitoring scheme with wearable devices and cloud service provider on encrypted medical data. Some other works [63–65, 71, 72] improve patients’ location privacy in mobile medical queries. Based on the Paillier encryption, Ayday et al. [73] introduce a privacy-preserving disease prediction scheme. However, due to its time-consuming exponential operations, the proposed scheme is not efficient in making the privacy-preserving comparison. Katzenbeisser et al. [74] introduce privacy-preserving recommendation systems, which use homomorphic public-key encryption schemes such as Paillier cryptosystem. The public-key encryption scheme is very time-consuming. Common symmetric encryption based systems (e.g. [75]) are very efficient, but key management becomes a challenging issue. We adopt data encryption techniques to fulfill privacy-preserving e-Healthcare schemes in this thesis. In chapter 3 and chapter 5, we compare our schemes with paillier-based schemes, and our schemes are much more efficient.

2.1.4 Differential privacy based schemes

Differential privacy is a tool used to offer quantifiable bound of privacy on the disclosure of query result. It is widely used in statistic analysis [13, 76]. Recent works [77–83] proposed

differential privacy based scheme on genomic analysis. They used these differential privacy mechanisms for detecting significant positions on DNA. Recently a paper [84] proposed a differential privacy scheme for privacy attack on Genomic Beacon Services [85]. However, because of the randomization added to the result, differential privacy based method is not suitable in some medical scenarios which require accurate results.

2.1.5 Disease Risk Detection

As early diagnosis of disease [86–89] can minimize the side-effects, safety risks, financial costs, etc, *disease risk prediction* has attracted the attention of medical and bioinformatics researchers. For example, in 2012, Anooj et al. [86] develop a fuzzy rule-based decision support system for the prediction of heart disease. In 2013, Bouwmeester et al. [90] use the multivariate logistic regression technique to develop a risk prediction model, in which a linear combination of predictors associated with multiple symptoms and environmental data are used to fit a logarithmic transformation of the probability of the tested disease. Disease risk detection is a topical research area, particularly in the use of big data analytics and ensuring that the privacy of user and health-related data is preserved. In this thesis, we adopt the risk prediction model from the work [90], then we propose some privacy-preserving techniques to fulfill the privacy-preserving e-Healthcare schemes.

2.2 Fundamental Concepts

2.2.1 Bilinear Pairings

Bilinear pairings [91] have been widely used to design protocols for key agreement, signatures and encryption. Let $\mathbb{G}, \mathbb{G}_T$ be two multiplicative cyclic groups with the same prime

order q . Suppose $\mathbb{G}$ and $\mathbb{G}_T$ are equipped with a pairing, i.e. a non-degenerated and efficiently computable bilinear map $e : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$, such that $e(g_1^a, g_2^b) = e(g_1, g_2)^{ab} \in \mathbb{G}_T$ for all $a, b \in \mathbb{Z}_q^*$, and any $g_1, g_2 \in \mathbb{G}$ in group $\mathbb{G}$, the Computational Diffie-Hellman (CDH) problem is hard. Given (g, g^a, g^b) for $g \in \mathbb{G}$ and unknown $a, b \in \mathbb{Z}_q^*$, it is intractable to compute g^{ab} in a polynomial time. However, the Decisional Diffie-Hellman (DDH) problem is easy. In other words, given (g, g^a, g^b, g^c) for $g \in \mathbb{G}$ and unknown $a, b, c \in \mathbb{Z}_q^*$, it is easy to determine whether $c = ab \pmod q$ by checking $e(g^a, g^b) \stackrel{?}{=} e(g^c, g)$.

Definition 1: A bilinear parameter generator gen is a probabilistic algorithm that takes a security parameter k as input, and outputs a 5-tuple $(q, g, \mathbb{G}, \mathbb{G}_T, e)$, where q is a k -bit prime number, and $e : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$ is a non-degenerated and efficiently computable bilinear map.

2.2.2 Bilinear Pairing with Composite Order

Bilinear groups of composite order [92] are groups with an efficient bilinear map where the group order is a product of two large primes, so that it supports homomorphic public key encryption. Let p, q be two large distinct prime numbers of the same κ -bit length, and set $N = pq$. $\mathbb{G}$ and $\mathbb{G}_T$ are two cyclic groups of the same composite order N . $\mathbb{G}$ and $\mathbb{G}_T$ are called bilinear map with composite order if there is a computable mapping $e : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$ with following properties:

- Bilinearity. $e(g_1^a, g_2^b) = e(g_1, g_2)^{ab}$ for any $a, b \in \mathbb{Z}_N$ and $g_1, g_2 \in \mathbb{G}$.
- Non-degeneracy. $e(g, g) \neq 1$.
- Computability. $e : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$ can be computed efficiently.

The definitions of composite bilinear generator and the subgroup decision problem are defined as below:

Definition 2. $\mathcal{Gen}$ is a probabilistic algorithm that takes a security parameter κ as input and output a 5-tuple $(g, N, \mathbb{G}, \mathbb{G}_T, e)$, where $N = pq$, p and q are two κ -bit length prime numbers.

Definition 3. Subgroup decision problem is shown as: Let g be a generator of $\mathbb{G}$, then $g_1 = g^p \in \mathbb{G}$ can generate the subgroup $\mathbb{G}_p = \{g_1^0, g_1^1, \dots, g_1^{p-1}\}$ of order p , and $g_2 = g^q \in \mathbb{G}$ can generate the subgroup $\mathbb{G}_q = \{g_2^0, g_2^1, \dots, g_2^{q-1}\}$ of order q . Given a tuple $(e, \mathbb{G}, \mathbb{G}_T, N, h)$, where h is drawn randomly from either $\mathbb{G}$ or the subgroup $\mathbb{G}_p$, decide whether $h \in \mathbb{G}_p$. The hard subgroup decision problem ensures the security of the BGN homomorphic cryptosystem defined below.

2.2.3 BGN Homomorphic Cryptosystem

The BGN cryptosystem was proposed by Boneh, Goh and Nissim [93], which is the first “somewhat homomorphic” cryptosystem with a constant-size ciphertext. The key idea in the BGN cryptosystem is based on the subgroup decision assumption, which supports polynomially many additions and just one multiplication. Concretely, it mainly contains three functions: key generation, encryption and decryption:

- $\mathcal{Gen}(\kappa)$: Given a secret parameter κ , choose two κ -bit prime number p, q and set $N = pq$. Let g be a generator of $\mathbb{G}$ with order N . Find a computable mapping $e : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$. Set $h = g^q$, which is a generator of the subgroup $\mathbb{G}$ with order p . Output the public key $pk = (N, \mathbb{G}, \mathbb{G}_T, e, g, h)$ and private key $sk = p$.
- $\mathcal{Enc}(pk, m)$: Given a message m from a small space, we choose a random value $r \in \mathbb{Z}_N$. Output the ciphertext $C = \mathcal{Enc}(pk, m) = g^m h^r \in \mathbb{G}$.

- $Dec(sk, C)$: Given a ciphertext $C \in \mathbb{G}$ and the secret key $sk = p$, perform the calculation $C^p = (g^m h^r)^p = (g^p)^m$. The message m is from small spaces, it suffices to solve the discrete log of $(g^p)^m$ with base g^p .

The BGN cryptosystem enjoys the following properties:

- Addition in $\mathbb{G}$: Given two ciphertext $Enc(m_1), Enc(m_2) \in \mathbb{G}$, we have $Enc(m_1) \cdot Enc(m_2) = Enc(m_1 + m_2)$.
- Addition in $\mathbb{G}_T$: Given two ciphertext $Enc_T(m_1), Enc_T(m_2) \in \mathbb{G}_T$, we have $Enc_T(m_1) \cdot Enc_T(m_2) = Enc_T(m_1 + m_2)$.
- Multiplication from $\mathbb{G}$ to $\mathbb{G}_T$: Given two ciphertext $Enc(m_1), Enc(m_2) \in \mathbb{G}$, we have $e(Enc(m_1), Enc(m_2)) = Enc_T(m_1 \cdot m_2) \in \mathbb{G}_T$.

2.2.4 Max heap

A max-heap is a complete binary tree, in which the value in each internal node is greater than or equal to the values in the children of that node [94]. The max-heap is widely used in top- k ranking algorithm [95]. Specifically, compared with the exhaustive element by element ranking, max heap is very efficient for finding the top- k elements when not need to know the order of this k elements. As shown in Fig. 2.1, mapping the elements of a max-heap into an array is trivial: If a node is stored at index k , then its left child is stored at index $2k + 1$ and its right child at index $2k + 2$, and its parent is at index $(k - 1)/2$, if exist.

Because a max-heap is a complete binary tree, it can be efficiently represented using a simple array. Moreover, given an array of N values, a heap containing those values can be built by simply shifting each internal node down to its proper location.

Figure 2.1: A sample of max heap

Cost of inserting a node into a max-heap with N nodes: same as the max-heap buildup, the inserting algorithm is involved in sifting nodes from the leaf to the root. The number of steps required for sifting values down will be maximized if the heap is full, which means $N = 2^d - 1$, and the heap height is d . The cost of inserting a node is $O(d)$, or $O(\lg N)$.

2.2.5 Pseudorandom Functions

A pseudorandom (PRF) is an efficiently computable keyed function $f_k(\cdot)$ whose value is indistinguishable, for a randomly chosen key k , from random elements in the function range. Let λ be the security parameter.

Definition 4: Algorithms F , on input $K \in \{0, 1\}^\lambda$, $x \in \{0, 1\}^*$, output $\{0, 1\}^\lambda$:

$$F : \{0, 1\}^\lambda \times \{0, 1\}^* \longrightarrow \{0, 1\}^\lambda$$

Definition 5: An algorithm F is a variable-input-length pseudorandom function if for all probabilistic polynomial-time function D :

$$Pr[D^{F(K_{\cdot})}(\lambda) = 1] - Pr[D^{f(\cdot)}(\lambda) = 1] \leq ngl(n)$$

where $ngl(n)$ is a negligible function in $K \in \{0, 1\}^\lambda$, f is a uniform choice of F_λ .

Definition 6: The oblivious PRF or OPRF is a protocol with a PRF $F(k, x)$ that a sender S inputs k to a receiver R with an input x , the R computes the value $F(k, x)$, but the sender S learns nothing.

2.2.6 IND-CCA2 Secure Public-Key Encryption technique

A public-key encryption technique is an asymmetric technique that uses a pairs of keys: a public key and a private key: a public key which can be distributed widely, and a private key which is only owned by the person who sets up the cyptographic system. A public-key encryption technique is a triplet of algorithms (Gen, Enc,Dec). Gen is a probabilistic polynomial-time key generation algorithm. Enc is a probabilistic polynomial-time encryption algorithm. Dec is a deterministic polynomial-time decryption algorithm.

Definition 7: To a two-stage adversary $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$ against a public-key encryption

technique PKE, we start the following experiment:

$$\begin{aligned}
 &Exp_{PKE, \mathcal{A}}^{cca2}(n) : \\
 &(pk, sk) \leftarrow Gen(1^n) \\
 &(m^0, m^1, state) \leftarrow \mathcal{A}_1^{Dec(sk, \cdot)}(pk) \text{ s.t. } |m^0| = |m^1| \\
 &b \leftarrow \{0, 1\} \\
 &c^* \leftarrow Enc(pk, m^b) \\
 &b' \leftarrow \mathcal{A}_2^{Dec(sk, \cdot)}(c^*, state) \\
 &\text{if } b = b' \text{ return } 1, \text{ else return } 0
 \end{aligned}$$

The adversary $\mathcal{A}_\epsilon$ is not allowed to query $Dec(sk, \cdot)$ with c^* . We define the advantage of $\mathcal{A}$ in the game as

$$Adv_{PKE, \mathcal{A}}^{cca2}(n) = |Pr[Exp_{PKE, \mathcal{A}}^{cca2}(n) = 1] - 1/2|$$

A public-key encryption (PKE) may be attacked by the adaptive chosen-ciphertext attacks. We say a PKE is indistinguishable against adaptive chosen-ciphertext attacks (IND-CCA2) if for all probabilistic polynomial-time(PPT) adversaries $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$ that make a polynomial number of oracle queries, the advantage of $\mathcal{A}$ in the experiment is a negligible function of n . For more comprehensive descriptions, please refer to [96].

2.2.7 Disease Risk Model

Many diagnosis prediction models combine patient characteristics and environmental data to predict the presence or absence of a certain disease [90]. The association between each symptom and a disease is expressed by the *odds ratio* (OR), which is the ratio of odds in a group of individuals having the symptom to that of those who do not have. The OR OR_i of a disease Y_i for some symptom predictors $A_i = \{a_1, a_2, \dots, a_m\}$, with each predictor value $a_j \in \{0, 1\}$ for $j = 1, 2, \dots, m$, is generally represented in terms of regression coefficients $B_i = \{b_1, b_2, \dots, b_m\}$ of the same length m . In other words, the regression coefficients work as the weights of the symptom predictors. In this way, the predicted risk of the disease Y_i with regards to the symptom A_i can be calculated as:

$$P(Y_i = 1|A_i) = \frac{1}{1 + \exp(-(\gamma + \sum_{j=1}^m a_j \cdot b_j))}, \quad (2.1)$$

where γ is an estimated intercept in the model. This model has been widely used in the medicine and clinician fields for disease risk tests and predictions [90]. To simplify the risk score calculation, the overall disease risk score S corresponding to the risk $P = P(Y_i = 1|A_i) = \frac{1}{1+\exp(-S)}$ can be computed by

$$S = \ln \frac{P}{1-P} = \gamma + \sum_{j=1}^m a_j \cdot b_j \quad (2.2)$$

For a complete description of the logistic regression model, we refer the interested reader to [90].

2.2.8 Determination of Disease Risk Threshold

In the above disease risk model, the regression coefficients $B_i = \{b_1, b_2, \dots, b_m\}$ and the estimated intercept γ for predicting some disease Y_i can be derived from the logistic regression model with a large volume of real-world medical data. In order to determine whether a user $U_l \in \mathbb{U}$ with the symptom predictors $A_i = \{a_1, a_2, \dots, a_m\}$ has the disease Y_i with a high probability, we can set a disease risk threshold S_{th} . If $\gamma + \sum_{j=1}^m a_j \cdot b_j \geq S_{\text{th}}$, then we can infer that user U_l has disease Y_i with a high probability. Otherwise, when $\gamma + \sum_{j=1}^m a_j \cdot b_j < S_{\text{th}}$, we infer that U_l has Y_i with a low probability. Because the disease risk model is an asset, the values $(B_i = \{b_1, b_2, \dots, b_m\}, \gamma, S_{\text{th}})$ should be kept private (i.e. privacy-preserving requirement). In the next section, we will present our PGuide scheme, which utilizes the disease risk model in an efficient and privacy-preserving way to achieve pre-clinical guidance for medical user.

Chapter 3

An Efficient and Privacy-Preserving Pre-Clinical Guide Scheme for Mobile e-Healthcare

3.1 Introduction

Little progress has been made in the commercial integration of processing clinical analytics while assuring the privacy of sensitive medical information [97]. For example, how can we be assured that our sensitive medical information are not been made available and exploited by third-party companies, such as insurance companies and future employers? In addition to the privacy requirements of patients, information leakage is a major concern for service providers as medical providers in countries such as U.S. are subject to exacting regulatory regime, e.g. Health Insurance Portability and Accountability Act (HIPAA).

In order to address the above-mentioned privacy challenges, and improve the accuracy

of disease risk prediction (and resulting in a shorter queue in hospitals), we propose an efficient privacy-preserving pre-clinical guidance service scheme (PGuide) designed to provide on-the-go medical guidance service while preserving user privacy. Using the PGuide scheme proposed in this chapter, the users can personally conduct pre-clinical diagnosis based on their health profiles [8] and obtain recommendation from trusted sources (e.g. hospitals and medical service providers) based on the diagnosis. In addition, the information transmitted to the hospitals and other medical service providers to calculate the disease risk use a disease prediction model in a privacy-preserving way. We then prove that our proposed PGuide scheme achieves the privacy-preservation for both the individual user and the medical service provider. Based on the findings from our evaluations, we show that our proposed PGuide scheme is efficient, in terms of computational cost and communication overhead.

The remainder of this chapter is organized as follows. In Section 3.2, we introduce our system model, security model and design goal. In Section 3.3, we introduce our pGuide scheme. The security analysis and performance evaluation are provided in Section 3.4 and Section 3.5, respectively. Section 3.6 concludes this chapter.

3.2 Models and Design Goal

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

3.2.1 System Model

In our system model, we focus on the disease risk calculation for the users with the help of one or more medical service providers. Our system model comprises four entities, namely:

Figure 3.1: System model under consideration

a group of m hospitals $H = \{H_1, H_2, \dots, H_m\}$, a healthcare centre as the trust authority (TA), a service provider (SP), and a number of users $\mathbb{U} = \{U_1, U_2, \dots, U_N\}$. The system model is illustrated in Fig. 3.1, where N, m indicate the numbers of medical users and hospitals, respectively.

- **Trust authority (TA):** Healthcare centre the trust authority, mainly responsible for initializing the system key materials for hospitals and medical users.
- **Hospitals $\mathbb{H} = \{H_1, H_2, \dots, H_m\}$:** Each hospital $H_j \in \mathbb{H}$ needs to have a one-off registration with the TA prior to processing user's disease queries, providing feedback to SP on whether it has the available resources (e.g. medical doctors) to treat the specific disease or not.
- **Service Provider (SP):** SP is the core entity, and is responsible for information processing, building the disease risk prediction model (see Section 2.2.7), and providing disease risk calculation service to users based on their health profiles. In addition,

SP directs disease queries from the users to the relevant hospitals, as well as recommending available hospitals to the users.

- Users $\mathbb{U} = \{U_1, U_2, \dots, U_N\}$: Each user $U_i \in \mathbb{U}$ has installed an app on their smart-phone(s). The app collects the user's health profile, sends them in a privacy-preserving way to the SP, and receives recommendations from the SP.

3.2.2 Security Model

In our security model, we assume that both the SP and users $\mathbb{U} = \{U_1, U_2, \dots, U_N\}$ are *honest but curious*. That is, SP will faithfully follow the disease risk diagnosis protocol, but may attempt to learn the users' health profile data. In addition, each U_i will not report false data, but may attempt to learn SP's disease risk predication model, which is regarded as SP's intellectual property (IP). We also assume that due to vested interest (e.g. reputation risk and criminal sanctions), a SP is not in collusion with the hospitals. Therefore, the following security requirements should be satisfied in the pre-clinical system.

- *Privacy Preservation.* Ensuring the privacy of user's sensitive health profile from SP is necessary, i.e. SP cannot learn the user's medical history and other sensitive information (e.g. blood type). In addition, information about user's diagnosis should not be learned by SP. It is also necessary to protect the intellectual property of the SP's disease prediction model (i.e. IP).
- *Authentication.* Authenticating the hospitals' recommendations to the users is important. For example, if a non-registered hospital (i.e. a hospital that has not been vetted as having appropriate standard) is recommended to the user, this could result in fatalities and subsequent law suits. Therefore, in the hospital recommendation system,

only hospitals that have registered with the TA can receive the user request.

We acknowledge that there are several other security attacks, such as forgery attacks and background attacks, in a typical disease risk calculation system. Since our focus is on privacy-preserving disease risk prediction, attacks not targeting privacy issue are beyond the scope of this chapter, and will be discussed in future work.

3.2.3 Design Goal

Our design goal is to develop a privacy-preserving pre-clinical guidance scheme to provide disease risk prediction and hospital recommendation. Using our scheme, a user seeking medical attention can have access to the disease risk predication services and hospital recommendation services from the medical service provider, without compromising on the privacy of user and SP. More specifically, the following two goals should be achieved.

- The proposed scheme should be effective, in terms of computation and communication. Despite the increase in computational resources available in smartphones, the storage and battery life are somewhat limited. Therefore, it is necessary to ensure that only lightweight computations are performed at the user-end and at the same time, do not overload the servers at SP or hospital.
- Security should be guaranteed. If the users are not assured that their health profiles are protected, then they will hesitate to (or not) use this service. Similarly, if the disease prediction model is not protected, then SP will not participate due to IP concerns. Therefore, the proposed scheme should also ensure that participating hospitals are authenticated.

3.3 Proposed PGuide Scheme

In this section, we propose our PGuide scheme. The scheme consists of two main phases, namely: a system setting and privacy-preserving pre-clinical guidance, together with its correctness analysis. The clinical guidance scheme has a privacy-preserving pre-disease diagnosis and a privacy-preserving hospital recommendation.

3.3.1 System Setting

The TA located at the healthcare centre will bootstrap the recommendation system. Specifically, given the security parameter k , the TA generates the bilinear parameters $(q, g, \mathbb{G}, \mathbb{G}_T, e)$ by running $gen(k)$ (see Section 2.2.1), and chooses a secure symmetric encryption algorithm $Enc()$ (i.e. AES) [98]. In addition, TA chooses two random numbers $(a, x) \in \mathbb{Z}_q^*$, as the master key, two random elements (h_1, h_2) in $\mathbb{G}$, and computes $b = H(a)$ ($H(\cdot)$ is a hash function), $A = g^a$, and $e(g, g)^x$. Finally, TA keeps the master key (a, b, x) secret, and publishes the system parameter $params = (q, g, \mathbb{G}, \mathbb{G}_T, e, h_1, h_2, A, e(g, g)^x, Enc())$.

When each hospital H_j registers itself with the healthcare centre, TA chooses two random numbers $(t_{j1}, t_{j2}) \in \mathbb{Z}_q^*$, and computes the access control key $ak_j = (g^{x+at_{j1}}, g^{t_{j1}}, g^{t_{j2}}, h_1^{t_{j1}} h_2^{t_{j2}})$ for the hospital H_j . The data acquisition process is crucial. For example, it was estimated that almost 80% of the time and effort is spent in cleaning and preparing real-world medical data before the data can be used by disease risk prediction model (see Section 2.2.7) [97]. Therefore, in the system setting phase of PGuide, a trustworthy data analytic company first communicates with hospitals to construct the disease risk model, and determines the regression coefficients $B_i = \{b_1, b_2, \dots, b_m\}$, the estimated intercept γ_i and the disease risk threshold $S_{th,i}$ (see Section 2.2.8) for each disease Y_i using some data mining methods [99, 100]. Each regression coefficient B_i is related to a disease predictor of a

3.3. PROPOSED PGUIDE SCHEME

disease Y_i . One disease has many related disease predictors. For example, the disease predictors for Parkinson’s Synucleinopathy-associated disease [8] include hyposmia, urinary dysfunction, specific sleep disturbances, depressive symptoms, and constipation.

According to each defined predictor $a_j \in A_i = \{a_1, a_2, \dots, a_m\}$ of a disease Y_i , a corresponding question $q_j \in Q_i = \{q_1, q_2, \dots, q_m\}$ is designed. For instance, as urinary dysfunction is a disease predictor of Parkinson’s disease, we design the question “*Do you have increased urinary frequency and urgency?*”, and the answer of each question q_i is in $a_j \in \{0, 1\}$. The result of the setup is shown in Table 3.1. For each disease, we obtain the corresponding regression coefficients, the estimated intercept, the disease risk threshold and the question set.

Table 3.1: Disease coefficients and the question set.

Disease	coefficients	γ	S_{th}	question set
Y_1	$\{b_{1,1}, b_{2,1}, \dots, b_{m,1}\}$	γ_1	$S_{th,1}$	$\{q_{1,1}, q_{2,1}, \dots, q_{m,1}\}$
Y_2	$\{b_{1,2}, b_{2,2}, \dots, b_{m,2}\}$	γ_2	$S_{th,2}$	$\{q_{1,2}, q_{2,2}, \dots, q_{m,2}\}$
Y_3	$\{b_{1,3}, b_{2,3}, \dots, b_{m,3}\}$	γ_3	$S_{th,3}$	$\{q_{1,3}, q_{2,3}, \dots, q_{m,3}\}$
...				
Y_n	$\{b_{1,n}, b_{2,n}, \dots, b_{m,n}\}$	γ_n	$S_{th,n}$	$\{q_{1,n}, q_{2,n}, \dots, q_{m,n}\}$

3.3.2 Privacy-preserving Pre-clinical Guidance

Pre-Disease Diagnosis. In the system setting, for each disease Y_i , we design a question set $q_j \in Q_i = \{q_1, q_2, \dots, q_m\}$, and the answer for each question q_i can be mapped to a binary value $a_i \in \{0, 1\}$. The general procedure of pre-clinical disease diagnosis can be described as follows:

- With the smartphone-based PGuide application, a user $U_l \in \mathbb{U}$ chooses a disease Y_i that he wishes to diagnose. A corresponding question set $Q_i = \{q_1, q_2, \dots, q_m\}$

will be shown in the application. After the U_l has answered all these questions Q_i , the answers representing U_l 's health profile will be mapped to a binary vector $A_i = \{a_1, a_2, \dots, a_m\}$ and transmitted to SP.

- Upon receiving A_i , SP will query the database to obtain the weighted coefficients $B_i = \{b_1, b_2, \dots, b_m\}$, the intercept γ_i and the threshold $S_{th,i}$ for the chosen disease Y_i . Then, SP will compute to check whether $A_i \cdot B_i + \gamma_i = \sum_{j=1}^m a_j \cdot b_j + \gamma_i \geq S_{th,i}$ and determine whether U_l has the disease Y_i with a high possibility. Finally, the result will be returned to U_l .
- Upon receiving the result, U_l will decide whether a doctor or a specialist needs to be consulted.

However, the above general pre-clinical procedure does not ensure the user's privacy. Therefore, in the following, we introduce a privacy-preserving comparison protocol (PPCP) in PGuide to ensure that another user is unable to learn the coefficients $B_i = \{b_1, b_2, \dots, b_m\}$, the intercept γ_i and the threshold $S_{th,i}$ in the risk model for the disease Y_i , and will not result in the disclosure of the user's health profile information $A_i = \{a_1, a_2, \dots, a_m\}$ to SP. The main steps of PPCP are summarized in Fig. 3.2, where $\vec{a} = A_i = \{a_1, a_2, \dots, a_m\} \in \mathbb{F}_2^m$ and $\vec{b} = B_i = \{b_1, b_2, \dots, b_m\} \in \mathbb{F}_q^m$. We remark that in the pre-clinical model, all b_i ($b_i > 0$), and each b_i is a small real number. For the efficient computation in PPCP, each b_i is increased by a factor of 10,000 times, such that all $\{b_1, b_2, \dots, b_m\}$ are integer values lying in $\mathbb{F}_q^m$ with $q = 2^{16}$. To ensure correctness and security of the proposed PPCP protocol, the security parameters $(l_0, l_1, l_2, l_3, l_4, l_5)$ are chosen first by the user U_l . The relationship between the PPCP parameters and these security parameters are shown in the Table 3.2.

The constraints of these security parameters are as follows: $l_1 < l_0$, $3l_2 + l_4 + l_5 < l_1$, $2l_2 + l_1 + l_3 + l_4 + l_5 < l_0 - 1$.

3.3. PROPOSED PGUIDE SCHEME

Table 3.2: Parameters setting for PPCP

PPCP Parameter	$ p $	$ \beta $	$ \alpha $	$ r_i $	$ q $	$ m $
Security Parameter	l_0	l_1	l_2	l_3	l_4	l_5

$|x|$ means the bit length of x .

Step 1 user chooses primes and random values: The user U_l chooses three large primes, α, β, p such that $|\alpha| = l_2$, $|\beta| = l_1$, $|p| = l_0$, a random number $s \in \mathbf{Z}_p^*$, and computes $s^{-1} \bmod p$.

Step 2 user calculates $\alpha, p, \vec{c}_i, \vec{c}_i'$: For each $a_i \in \vec{a}$, three random numbers (x_i, y_i, r_i) are chosen with the constraint $x_i + y_i = r_i \cdot \beta$, $\frac{r_i \cdot \beta}{2} < y_i < r_i \cdot \beta$ and $|r_i| = l_3$. U_l computes the vectors

User: $(a_1, a_2, \dots, a_m) \in \mathbf{F}_2^m$	SP: $(b_1, b_2, \dots, b_m) \in \mathbf{F}_q^m, \gamma, S_{th}$
1. choose large primes α, β, p such that $ \alpha = l_2$, $ \beta = l_1$, $ p = l_0$, and a large random $s \in \mathbf{Z}_p^*$; 2. for $i = 1, 2, \dots, m$, choose random numbers x_i, y_i, r_i , such that $x_i + y_i = r_i \cdot \beta$; compute $c_i = \alpha \cdot a_i + x_i$, $c_i' = s \cdot y_i \bmod p$ end for choose a random y_0 such that $ y_0 < \alpha $, compute $c_0' = s \cdot y_0 \bmod p$	
$\xrightarrow{\alpha, p, \vec{c}_i, \vec{c}_i'}$ 3. choose random numbers t_1, t_2, t_3 such that $ t_2 < \alpha $, $ t_1 \cdot t_2 > \alpha^2 $, $ t_3 < \alpha $, and $ t_1 + t_3 \cdot \alpha < \alpha^2 $ for $i = 1, 2, \dots, m$: $D_i = \alpha \cdot b_i \cdot c_i, D_i' = \alpha \cdot b_i \cdot c_i' \bmod p$ end for $D_0' = t_3 \cdot c_0' \bmod p, D = t_2 \cdot (\sum_{i=1}^m D_i + \gamma \cdot \alpha^2 - S_{th} \cdot \alpha^2 + t_1)$ $D' = t_2 \cdot (\sum_{i=1}^m D_i' + D_0') \bmod p$	
$\xleftarrow{D, D'}$ 4. $E' = s^{-1} \cdot D' \bmod p$, $E = (D + E') \bmod \beta$ if $ E = \beta $, then $\vec{a} \cdot \vec{b} + \gamma < S_{th}$; else : $\vec{a} \cdot \vec{b} + \gamma \geq S_{th}$	

Figure 3.2: PPCP protocol

$\vec{c} = \{c_1, c_2, \dots, c_m\}$, $\vec{c}' = \{c_0', c_1', c_2', \dots, c_m'\}$, where each (c_i, c_i') is

$$\begin{aligned} c_i &= \alpha \cdot a_i + x_i, & c_i' &= s \cdot y_i \bmod p, \text{ for } i = 1, 2, \dots, m \\ c_0' &= s \cdot y_0 \bmod p, \text{ where } y_0 < \alpha \text{ is a random number} \end{aligned} \quad (3.1)$$

and sends $(\alpha, p, \vec{c}_i, \vec{c}_i')$ to SP. Because of the large prime α , the random numbers x_i, y_i and $\bmod p$ operation, SP is unable to determine whether $a_i \in \vec{a}$ is 1 or 0.

Step 3 SP calculates D, D' : After receiving $(\alpha, p, \vec{c}_i, \vec{c}_i')$, SP chooses three random numbers t_1, t_2, t_3 with the constraints $|t_2| < |\alpha|$, $|t_3| < |\alpha|$, $|t_1 \cdot t_2| > |\alpha^2|$, $|t_1 + t_3 \cdot \alpha| < |\alpha^2|$, and computes the vectors $\vec{D}, \vec{D}'$, where

$$\begin{aligned} D_i &= \alpha \cdot b_i \cdot c_i, D_i' = \alpha \cdot b_i \cdot c_i' \bmod p, \text{ for } i = 1, 2, \dots, m \\ D_0' &= t_3 \cdot c_0' \bmod p \end{aligned} \quad (3.2)$$

Then, SP computes

$$\begin{aligned} D &= t_2 \cdot \left(\sum_{i=1}^m D_i + \gamma \cdot \alpha^2 - S_{\text{th}} \cdot \alpha^2 + t_1 \right) \\ D' &= t_2 \cdot \left(\sum_{i=1}^m D_i' + D_0' \right) \end{aligned} \quad (3.3)$$

SP will now return (D, D') to U_l , without revealing the values $b_i \in \vec{b}_i$ and the threshold of the disease risk S_{th} to U_l .

Step 4 user calculates whether $\vec{a} \cdot \vec{b} + \gamma < S_{\text{th}}$: Upon receiving the data (D, D') , U_l first computes

$$E' = s^{-1} \cdot D' \bmod p, \quad E = (D + E') \bmod \beta \quad (3.4)$$

Finally, U_l can determine whether $\vec{a} \cdot \vec{b} + \gamma < S_{\text{th}}$ from the bit length of E . If $|E| = |\beta|$, then

3.3. PROPOSED PGUIDE SCHEME

U_l can determine $\vec{d} \cdot \vec{b} + \gamma < S_{\text{th}}$. Otherwise, $\vec{d} \cdot \vec{b} + \gamma \geq S_{\text{th}}$.

Correctness. The correctness of PPCP can be illustrated as follows: Given the constraints of the security parameters shown above, we obtain:

$$\begin{aligned} 2l_2 + l_1 + l_3 + l_4 + l_5 &< l_0 - 1 \\ \Rightarrow \alpha^2 \cdot m \cdot q \cdot r_i \cdot \beta &< p/2 \end{aligned} \quad (3.5)$$

In addition, the relationship of α and p can be calculated as follows:

$$\begin{aligned} 3l_2 + l_4 + l_5 &< l_1, \quad l_1 < l_0 \\ \Rightarrow 3l_2 + l_4 + l_5 &< l_0 - 1 \Rightarrow 3l_2 < l_0 - 1 \\ \Rightarrow \alpha^3 &< p/2 \end{aligned} \quad (3.6)$$

In step 3, SP receives the data and calculates:

$$\begin{aligned} D_i &= \alpha \cdot b_i \cdot c_i = \alpha \cdot b_i \cdot (\alpha \cdot a_i + x_i) = \alpha^2 \cdot a_i \cdot b_i + \alpha \cdot b_i \cdot x_i \\ D_i' &= \alpha \cdot b_i \cdot c_i' = \alpha \cdot b_i \cdot s \cdot y_i \bmod p, \text{ for } i = 1, 2, \dots, m \\ D_0' &= t_3 \cdot c_0' = t_3 \cdot s \cdot y_0 \bmod p \end{aligned} \quad (3.7)$$

Then, SP can compute (D, D') , where

$$\begin{aligned} D &= t_2 \cdot \left(\alpha^2 \cdot \sum_{i=1}^m a_i \cdot b_i + \alpha \cdot \sum_{i=1}^m b_i \cdot x_i + \alpha^2 \cdot \gamma - \alpha^2 \cdot S_{\text{th}} + t_1 \right) \\ D' &= t_2 \cdot \left(\alpha \cdot \sum_{i=1}^m b_i \cdot s \cdot y_i + t_3 \cdot s \cdot y_0 \right) \bmod p \end{aligned} \quad (3.8)$$

In step 4, U_l removes the factor s from D' by multiplying $s^{-1} \bmod p$:

$$\begin{aligned}
 E' &= s^{-1} \cdot D' = s^{-1} t_2 \cdot \left(\alpha \cdot \sum_{i=1}^m b_i \cdot s \cdot y_i + t_3 \cdot s \cdot y_0 \right) \bmod p \\
 &= t_2 \cdot \left(\alpha \cdot \sum_{i=1}^m b_i \cdot y_i + t_3 \cdot y_0 \right) \bmod p \\
 &\xrightarrow{\because (t_2 \cdot \alpha \cdot \sum_{i=1}^m b_i \cdot y_i) < (\alpha^2 \cdot m \cdot q \cdot r_i \cdot \beta) < p/2 \quad \text{Eq.(3.5)}} \\
 &\xrightarrow{\because \text{and} \quad (t_2 \cdot t_3 \cdot y_0) < \alpha^3 < p/2 \quad \text{Eq.(3.6)}} \\
 &= t_2 \cdot \left(\alpha \cdot \sum_{i=1}^m b_i \cdot y_i + t_3 \cdot y_0 \right)
 \end{aligned} \tag{3.9}$$

In the last calculation, U_l obtains

$$\begin{aligned}
 E = D + E' &= t_2 \cdot \left(\alpha^2 \cdot \sum_{i=1}^m a_i \cdot b_i + \alpha \cdot \sum_{i=1}^m b_i \cdot x_i \right. \\
 &\quad \left. + \alpha^2 \cdot \gamma - \alpha^2 \cdot S_{\text{th}} + t_1 + \alpha \cdot \sum_{i=1}^m b_i \cdot y_i + t_3 \cdot y_0 \right) \bmod \beta \\
 &= t_2 \cdot \left[\alpha^2 \cdot \left(\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}} \right) \right. \\
 &\quad \left. + \alpha \cdot \sum_{i=1}^m b_i \cdot \beta + t_1 + t_3 \cdot y_0 \right] \bmod \beta
 \end{aligned} \tag{3.10}$$

Let $k_{\text{gap}} = |\beta| - |\alpha^2| - |q| - |t_2|$, and $k_{\text{gap}} > 200$. For example, $|\alpha| = 160$, $|\beta| = 700$, $|p| = 1024$,

$|t_1|= 300, |t_2|= 100, |t_3|= 100, |q|= 16$. If $\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}} \geq 0$, Eq. (3.10) becomes

$$\begin{aligned}
 E &= t_2 \cdot \left[\alpha^2 \cdot \left(\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}} \right) + t_1 + t_3 \cdot y_0 \right] \bmod \beta \\
 &\xrightarrow{\because k_{\text{gap}} = |\beta| - |\alpha^2| - |q| - |t_2|, k_{\text{gap}} > 200, t_2 \cdot (t_1 + t_3 \cdot y_0) < \alpha^3 < \beta \text{ and } |\alpha^3| < |\beta|} \\
 &= t_2 \cdot \left[\alpha^2 \cdot \left(\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}} \right) + t_1 + t_3 \cdot y_0 \right]
 \end{aligned} \tag{3.11}$$

Because $|t_1 + t_3 \cdot \alpha| < |\alpha^2|$, the length of E is dominated by $t_2 \cdot \alpha^2 \cdot (\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}})$, that is, $|E| \approx |\alpha^2| + |q| + |t_2| \ll |\beta|$. It is easy to observe that the bit length of E is much less than that of β when $\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}} \geq 0$.

On the other hand, if $\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}} < 0$, Eq. (3.10) becomes

$$\begin{aligned}
 E &= t_2 \cdot \alpha^2 \cdot \left(\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}} \right) + \beta \\
 &\quad + \left(t_2 \cdot \alpha \cdot \sum_{i=1}^m b_i - 1 \right) \cdot \beta + t_2 \cdot (t_1 + t_3 \cdot y_0) \bmod \beta \\
 &= t_2 \cdot \alpha^2 \cdot \left(\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}} \right) + \beta \\
 &\quad + t_2 \cdot (t_1 + t_3 \cdot y_0) \bmod \beta \\
 &\xrightarrow{\because t_2 \cdot \alpha^2 \cdot (\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}}) + \beta < \beta, t_2 \cdot (t_1 + t_3 \cdot y_0) < \alpha^3 < \beta \text{ and } |\alpha^3| < |\beta|} \\
 &= t_2 \cdot \alpha^2 \cdot \left(\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}} \right) + \beta + t_2 \cdot (t_1 + t_3 \cdot y_0)
 \end{aligned} \tag{3.12}$$

Because $t_2 \cdot (t_1 + t_3 \cdot y_0) < \alpha^3 < \beta$, $t_2 \cdot \alpha^2 \cdot (\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}}) < 0$ and $|t_2 \cdot \alpha^2 \cdot (\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}})| \ll |\beta|$, we have the length of E dominated by β , that is, $|E| = |\beta|$.

From the above observations, U_l can distinguish the disease risk from the bit length of E . Thus, the correctness of the PPCP protocol is satisfied.

Recommendation of hospitals. After diagnosing several diseases $\mathcal{Y} = \{y_1, y_2, \dots, y_w\}$ using the proposed PPCP protocol, the user U_l can choose a high-risk disease $y_i \in \mathcal{Y}$ that he/she wishes to seek treatment for. SP will query all hospitals within U_l 's district or vicinity whether these hospitals have medical doctors available to treat this disease without the need for SP to know the specific disease. Then SP forwards the encrypted response from hospitals to U_l . Note that the number of the hospitals within U_l 's vicinity should not be large. It does not take much time for U_l to decrypt these responses and determine which hospital(s) had replied with a “yes”. It is trivial to note that the privacy-preserving recommendation protocol helps to ensure that the privacy of the user's disease / treatment sought is preserved. In addition, only authenticated hospitals are recommended to the user.

Step 1: The app on U_l 's smartphone first chooses a random number $s \in \mathbb{Z}_q^*$, computes $sk = e(g, g)^{xs}$ and $c = (c_1, c_2, c_3)$ as

$$c_1 = g^s, c_2 = A^s \cdot h_1^{-s}, c_3 = h_2^{-s} \quad (3.13)$$

U_l sends $Enc_{sk}(y_i), c = (c_1, c_2, c_3)$ and the current location to SP. SP chooses hospitals within the user's district or vicinity and sends them the encrypted message $(Enc_{sk}(y_i), c_1, c_2, c_3)$. Note that, in this scenario, the location privacy preservation for user is currently not considered, as it is not as critical as the user's sensitive health information.

Step 2: Upon receiving $(Enc_{sk}(y_i), c_1, c_2, c_3)$, each hospital H_j will perform the following steps:

- Uses the access control key $ak_j = (g^{x+at_{j1}}, g^{t_{j1}}, g^{t_{j2}}, h_1^{t_{j1}} \cdot h_2^{t_{j2}})$ to compute

$$\begin{aligned}
 sk &= \frac{e(c_1, g^{x+at_{j1}})}{e(g^{t_{j1}}, c_2) \cdot e(g^{t_{j2}}, c_3) \cdot e(h_1^{t_{j1}} h_2^{t_{j2}}, c_1)} \\
 &= \frac{e(g^s, g^x g^{at_{j1}})}{e(g^{g_{j1}}, g^{as} \cdot h_1^{-s}) \cdot e(g^{t_{j2}}, h_2^{-s}) \cdot e(h_1^{t_{j1}} h_2^{t_{j2}}, g^s)} \\
 &= \frac{e(g^s, g^x) e(g^s, g^{at_{j1}})}{e(g^{t_{j1}}, g^{as}) e(g^{t_{j1}}, h_1^{-s}) e(g^{t_{j2}}, h_2^{-s}) e(h_1^{t_{j1}} h_2^{t_{j2}}, g^s)} \\
 &= \frac{e(g^s, g^x)}{e(g^s, h_1^{t_{j1}} h_2^{t_{j2}})^{-1} e(h_1^{t_{j1}} h_2^{t_{j2}}, g^s)} = e(g, g)^{xs}
 \end{aligned} \tag{3.14}$$

- Computes $y_i = Dec_{sk}(Enc_{sk}(y_i))$. The hospital information system (HIS) in each hospital H_j will return $Enc_{sk}(yes|timestamp)$ to SP if they have the medical doctor available to treat the patient y_i , otherwise they will return $Enc_{sk}(no|timestamp)$
- SP collects the encrypted responses from the participating hospitals and forwards them to U_l . As the number of the hospitals within U_l 's vicinity is not large, it will not require U_l to spend much time in decrypting the responses and determining which hospitals had responded with a “yes”.

The hospital recommendation system will reduce the waiting time due to a mismatch or going to a hospital that does not have an appropriate or available medical doctor to treat the patient.

3.4 Security Analysis

In this section, we analyze the security of our proposed PGuide scheme, particularly focusing on how the proposed PPCP protocol can achieve the privacy-preservation of a user's health profile ($A = \vec{a} = \{a_1, a_2, \dots, a_m\}$) and the SP's IP ($B = \vec{b} = \{b_1, b_2, \dots, b_m\}, \gamma, S_{th}$) in

the disease diagnosis phase. We will also examine the privacy-preservation of the disease information in the hospital recommendation phase.

• **Security of health profile in user query.** In a user query in PGuide, sensitive health profile information is encrypted by the PPCP scheme. A query consists of two vectors $\vec{c} = (c_1, \dots, c_m)$, $\vec{c}' = (c'_0, c'_1, \dots, c'_m)$ and two large prime numbers α, p . As SP is honest but curious, it may attempt to recover $\vec{a} = (a_1, \dots, a_m)$ using exhaustive attacks on $\vec{c}$ and $\vec{c}'$. However, the components $\vec{c}$ and $\vec{c}'$ can be viewed as an equation group of $2m$ equations with $4m + 1$ unknowns $s, (x_i, y_i, a_i, w_i)$, for $i = 1, 2, \dots, m$, as below

$$\left\{ \begin{array}{l} c_1 = \alpha \cdot a_1 + x_1 \\ \vdots \\ c_m = \alpha \cdot a_m + x_m \\ c'_1 = s \cdot y_1 \bmod p \\ \vdots \\ c'_m = s \cdot y_m \bmod p \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} c_1 = \alpha \cdot a_1 + x_1 \\ \vdots \\ c_m = \alpha \cdot a_m + x_m \\ c'_1 = s \cdot y_1 + w_1 \cdot p, w_1 \in \mathbf{Z}_{\geq 0} \\ \vdots \\ c'_m = s \cdot y_m + w_m \cdot p, w_m \in \mathbf{Z}_{\geq 0} \end{array} \right.$$

Because the number of unknowns (i.e. $4m + 1$) is more than those in the equations (i.e. $2m$), this equation group is not determined. That is, SP is unable to learn $\vec{a}$ by solving this equation group.

• **Security of SP's disease risk model.** The user is also honest but curious, and may seek to recover the coefficients of disease risk model and the threshold by generating and solving an over-determined polynomial equation group. If we do not include the random numbers t_1, t_2, t_3 , the user may reveal the disease risk model ($\vec{b} = \{b_1, b_2, \dots, b_m\}, \gamma, S_{th}$) as follows.

i) A user may attempt to reveal the disease risk model by attacking E . Without t_1, t_2, t_3 ,

after issuing a query with $\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}} > 0$, the user can obtain the encrypted value

$$E = \alpha^2 \cdot \left(\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}} \right) + \alpha \cdot \sum_{i=1}^m b_i \cdot \beta \bmod \beta$$

$$\xrightarrow{\because |\alpha^2 \cdot (\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}})| < |\beta|}$$

$$= \alpha^2 \cdot \left(\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}} \right)$$

The unknowns in E are the coefficients $\{b_1, b_2, \dots, b_m\}$, the intercept γ and the threshold S_{th} of the disease model. After issuing k queries with k different $\vec{a}$, a group of k equations and $m + 2$ unknowns $\{b_1, b_2, \dots, b_m\}, \gamma, S_{\text{th}}$ can be generated.

$$\begin{cases} E_1 = \alpha^2 \cdot (\sum_{i=1}^m a_{i,1} \cdot b_i + \gamma - S_{\text{th}}) \\ \vdots \\ E_k = \alpha^2 \cdot (\sum_{i=1}^m a_{i,k} \cdot b_i + \gamma - S_{\text{th}}) \end{cases}$$

Once k is more than $m + 2$, the number of unknowns is less than those in the equations. Consequently, this equation group is over-determined and the disease risk model can be revealed. In order to prevent such an attack, we configure two random numbers t_1 and t_2 . For each user query, t_2 is a distinct random number, which increases the number of unknowns linearly with the number of queries. Thus, the equation group is not determined. However, if we only configure t_2 without t_1 , a user may reveal $\alpha^2 \cdot (\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}})$ based on two queries with the same vector $\vec{a}$, i.e., the common divisor of $E_j = t_{2,j} \cdot \alpha^2 \cdot (\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}})$, for $j = 1, 2$. After obtaining $m + 2$ components $\alpha^2 \cdot (\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}})$, for $j = 1, 2, \dots, m + 2$, a user can also reveal the disease risk model as above. Therefore, to deal with this vulnerability, we add the random number t_1 .

Then, with two queries on the same $\vec{d}$, a user can get

$$E_j = t_{2,j} \cdot \alpha^2 \cdot \left(\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}} \right) + t_{2,j} \cdot t_{1,j}, j = 1, 2$$

If $|t_{1,j} \cdot t_{2,j}| < |\alpha^2|$, then $t_{2,j} \cdot \alpha^2 \cdot (\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}})$ can be derived from $E_j - E_j \bmod \alpha^2$.

Then again, the disease risk model can be revealed using the above attack. Therefore, we need the constraint $|t_1 \cdot t_2| > |\alpha^2|$ to prevent such an attack.

ii) A user may also attempt to reveal the disease risk model by attacking E' . If we configure t_1, t_2 without t_3 , in step 4, the user can obtain the value

$$E' = t_2 \cdot \alpha \cdot \left(\sum_{i=1}^m b_i \cdot y_i \right) \bmod p$$

Then, the user can reveal $\alpha \cdot (\sum_{i=1}^m b_i \cdot y_i)$ from the two queries issued to the same vector $\vec{y} = (y_1, y_2, \dots, y_m)$ (i.e. the common divisor of $E'_j = t_{2,j} \cdot \alpha \cdot (\sum_{i=1}^m b_i \cdot y_i)$, $j = 1, 2$).

To prevent such an attack, we configure a random number t_3 on $c_0' = s_0 \cdot y_0 \bmod p$. Then, in step 4, the user obtains

$$E' = t_2 \cdot \alpha \cdot \left(\sum_{i=1}^m b_i \cdot y_i \right) + t_2 \cdot t_3 \cdot y_0 \bmod p$$

In addition, the constraint $|t_1 + t_3 \cdot \alpha| < |\alpha^2|$ is necessary, which prevents the random numbers from changing the result in E .

• **Security of disease information and authentication of hospitals in hospital recommendation protocol.** It is easy to see that $e(g, g)^{xs}$ can be recovered only by a registered hospital $H_j \in \mathbb{H}$ with its access key $ak_j = (g^{x+at_{j1}}, g^{t_{j1}}, g^{t_{j2}}, h_1^{t_{j1}} h_2^{t_{j2}})$ from $(c_1 = g^s, c_2 = A^s \cdot h_1^{-s}, c_3 = h_2^{-s})$, and the information about the disease y_i can only be obtained

using the appropriate symmetric key $e(g, g)^{xs}$ and $Dec()$. There are many common diseases can be treated in different hospitals [101], and SP only receives encrypted feedback $Enc_{sk}(yes|timestamp)$ or $Enc_{sk}(no|timestamp)$ from the participating hospitals. Therefore, SP will not be able to learn the details of the disease. In addition, only the hospitals who have registered with TA will have the access control key $ak_j = (g^{x+at_{j1}}, g^{t_{j1}}, g^{t_{j2}}, h_1^{t_{j1}} \cdot h_2^{t_{j2}})$ to decrypt the encrypted user request. Such an authentication scheme prevents non-registered and non-vetted hospitals from participating.

Based on the above security analysis, we have shown that our proposed PGuide scheme provides privacy preservation for user health profile, disease and SP's disease risk prediction model, assuming that there is no collusion between hospitals and SP.

3.5 Performance Evaluation

In this section, we evaluate our proposed PGuide scheme in two stages (i.e. Pre-disease Diagnosis; and Recommendation of Hospitals), in terms of computational cost and communication overhead.

3.5.1 Pre-disease Diagnosis

Experimental setup. We design a PGuide Android app, as shown in Fig. 3.3, in addition to a server side application and a hospital side application on the Tomcat Apache server 8. The detailed experimental settings are outlined in Table 3.3. In order to demonstrate the efficiency of PPCP in Pre-disease Diagnosis, we also develop a scheme with the same function of PPCP but based on typical Paillier encryption with modulus $|n^2| = 2048$ [1] as a baseline scheme for comparison. We choose the length of vectors $\vec{a}, \vec{b}$ as

Table 3.3: Experiment setup

(a) Testbed setting

Role	Machine	Hardware & Software
SP	PC	3.1. GHz processor, 8GB RAM and Window 7 platform
Medical user	MI-ONE Plus	Android 4.1.2 system, dual core 1.5 GHz processor, and 1 GB RAM
Hospital	Mac Pro	2.9 GHz processor, 8GB RAM and OSX Yosemite platform

(b) Parameter setting for pre-disease diagnosis

Parameter	$ \alpha $	$ \beta $	$ p $	$ t_1 $	$ t_2 $	$ t_3 $	$ q $	$ r_i $
Setting	160	700	1024	300	100	100	16	100

(c) Parameter setting for hospital recommendation

Parameter	Curve	q	$ r $	$ q $
Setting	$y^2 = x^3 + x$	$q = 3 \pmod{4}$	160	512

Figure 3.3: Prototype: Android app for Pre-disease Diagnosis

3.5. PERFORMANCE EVALUATION

$m = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ in the experiments, and run the experiments 100 times. The average results of these 100 experiments are reported below.

Computational cost for Pre-disease Diagnosis Fig. 3.4(a) and Fig. 3.4(b) plot the computational costs with varying vector lengths from 10 to 100. From Fig. 3.4(a), it is clear that our proposed PGuide scheme is significantly faster than the Paillier-based scheme. As our proposed PGuide scheme does not employ computationally expensive exponent operations, user's cost requirement is low, and significantly lower than the computational cost in the Paillier-based scheme. Given the m -length vector, the total operation count in user side is $3m+3$, so the time complexity of the user side is $O(m)$. As shown in Fig. 3.4(b), the server's computational costs in the Paillier-based scheme are low with the vector length m because all the disease coefficients b_i are small integers and the exponent operation of ciphertext $Enc(a_i)^{b_i}$ does not require much time. Nonetheless, the computational cost of our proposed PGuide scheme at the server-end is still significantly less than that of the Paillier-based scheme at the server-end, as we do not require any exponent operations. The total operation count at the server-end is $4m + 11$, so the time complexity at the server-end is $O(m)$.

Communication overhead for Pre-disease Diagnosis Fig. 3.4(c) plots the communication overhead with varying vector lengths (i.e. m from 10 to 100). Based on the above parameter settings, the length of $\vec{c}$ is the same as the length of $(\beta + x_i) \cdot m$, which is $(700 + 100)m$. The length of $\vec{c}'$ is at most $m \cdot p$, which is $1024m$. The lengths of α and p are $|\alpha| = 160$ and $|p| = 1024$. Therefore, the length of the ciphertext $(\vec{c}, \vec{c}', \alpha, p)$ in the proposed PGuide scheme is at most $1184 + 1824m$ bits. On the other hand, the length of m ciphertext of Paillier-based scheme is the same as $n^2 \cdot m$ ($2048m$), which is larger than that of our proposed PGuide scheme.

3.5. PERFORMANCE EVALUATION

(a) Computational time at user-end with different vector lengths (b) Computational time at server-end with different vector lengths (c) Length of ciphertext with different vector lengths

Figure 3.4: Latency in the pre-disease diagnosis process for our proposed pGuide and paillier-based scheme [1]

3.5.2 Recommendation of Hospitals

(a) Hospital recommendation Android app (b) Average response time (c) Throughput

Figure 3.5: Recommendation of hospitals Android app

Experimental setup. In our Hospital Recommendation Android app (see Fig. 3.5(a)), the user can choose a disease in the disease list page as well as obtaining the corresponding recommended hospitals. We adopt the JPBC library [102] to implement the underlying pairing algorithm, using the pairing parameters shown in Table 3.3(c). In addition, a 128-bit symmetric encryption algorithm is used to encrypt the user request. We used the SHA-256 hash function to hash the pairing element $e(g, g)^{xs}$ to a 256-bit length key, which is truncated to the 128-bit length symmetric encryption key.

In our experiments, the pairing parameters were generated and stored in a file. Android users store this parameter file in the SD-card, and the hospital-side application stores the file on a local disk. At the user-end, it is not costly to compute the symmetric key sk and $c = (c_1, c_2, c_3)$ because the users can conduct a one-off calculation of the keys offline. The time-consuming processes are the computations of the symmetric key sk with $c = (c_1, c_2, c_3)$ and the access control key ak at the hospital-side because of the bilinear pairing calculations. We developed a common AES encryption as a reference for comparison. For a fair comparison of response time and throughput, we used the Apache JMeter [103] to simultaneously send n user requests. $n = \{100, 200, \dots, 1000\}$. The results are reported below.

Response time and throughput of Hospital Recommendation Fig. 3.5(b) and Fig. 3.5(c) show the average response time and throughput with a varying number of user requests. The average response time of our proposed single-attribute encryption protocol is nearly twice of that of common symmetric encryption-based protocol. This is because our hospital recommendation protocol is based on AES and Bilinear Pairing encryption, which is more time-consuming than common symmetric encryption with a given key. The gaps illustrated in Fig. 3.5(b) and Fig. 3.5(c) are acceptable for practical applications, and our protocol offers more convenience in terms of key management. Because in our protocol, the hospital can easily calculate the secret key $sk = e(g, g)^{xs}$ with $c = (c_1, c_2, c_3)$, while in common symmetric encryption protocol, the hospital needs to store all the different secret keys from different users.

Fig. 3.6(a) shows that the message encryption time at the user-side grows linearly over the number of hospitals for common symmetric encryption based protocol while that of our proposed single-attribute encryption protocol remains constant. This is because the

common symmetric encryption based protocol generates as many messages as the number of hospitals, while our proposed protocol performs a one-off offline calculation of $c = (c_1, c_2, c_3)$, and encrypts the message online only once in every request.

Fig. 3.6(b) illustrates that the key size of user-side grows linearly with the number of hospitals for the common AES symmetric encryption based protocol. The large number of keys could be leaked if the users are required to maintain the keys of all hospitals. Our proposed protocol solves this key management issue by computing the secret key $sk = e(g, g)^{xs}$ with $c = (c_1, c_2, c_3)$ at hospital side.

Figure 3.6: Comparative Summary: Key generation time, key sizes for user-side and communication cost with varying number of hospitals

Communication cost of Hospital Recommendation Fig. 3.6(c) compares the computational cost between our proposed protocol and the common symmetric encryption based protocol. In our protocol, the length of $\mathbb{G}$ is 128 bytes, and we assume that the length of the disease name Y_i is 24 bytes, then the length of the payload for medical user request $c_1, c_2, c_3, Enc_{sk}(Y_i)$ is $128 * 3 + 24$ bytes. On the other hand, the common symmetric encryption based protocol has linearly increasing message size because the user has to encrypt as many pieces of disease information as the number of hospitals just for one recommendation. Therefore, our proposed protocol is clearly more efficient and scalable.

3.6 Conclusions

In this chapter, we have proposed an efficient and privacy-preserving pre-clinical guidance scheme (PGuide). The scheme has two key phases. Firstly, it employs a privacy-preserving comparison protocol (PPCP), which enables a user to obtain disease risk predication services from a service provider without compromising the privacy of the user and the server provider. Secondly, it employs a single-attribute encryption technique to conduct an efficient privacy-preserving hospital recommendation service. Our security analysis demonstrated that the PPCP and hospital recommendation service achieve the privacy-preserving requirements. Evaluations demonstrated the efficiency and scalability of our scheme for real-world deployment. Specifically, at the user side, our proposed PGuide scheme's computational latency is at most 10 ms, which is significantly smaller than the widely used paillier-based scheme's computational latency, which climbs from 120 ms to 1200 ms for the length of vector m range from 10 to 100. Moreover, at the server side, our proposed PGuide scheme is around 20 times faster than the paillier-based scheme.

Chapter 4

Efficient and Privacy-Preserving Health Query over Outsourced Cloud

In the privacy-preserving comparison protocol in last chapter, we have discussed the privacy-preserving healthcare query in a two-party system model, which only includes the users and the healthcare service provider. In this chapter, we will discuss a three-party system model, in which, the healthcare center outsources the computation tasks to the cloud server, and the users conduct healthcare query from the cloud.

4.1 Introduction

A few years ago, the research on wireless sensor network technologies and applications led to the introduction of Body Sensor Networks (BSNs). With the pervasiveness of smart phones and BSNs, health query service has received considerable attentions and become more popular. Particularly, a user is equipped with a wearable BSN, which comprises wireless physiological sensors like, smart fabrics, skin electrodes, surface thermistor or

4.1. INTRODUCTION

three-axis accelerometer for motion sensing. Various BSN applications are proposed in these years including early detection, elderly assistance, physical activity monitoring and so on [17–19]. Smart devices like smart phones and smart watches are used to collect these sensor data stream and conduct some signal processing and algorithms [10–12]. As shown in Fig.4.1, the benefits of the wearable health system is the possibility to enjoy the health monitor service without obstructing user’s comfort in performing daily activities.

Figure 4.1: Wearable health monitor system

While the BSN health query service is popular and vital, most of the health query service require users to submit physiological data, medical history, personal information, which raises concerns about leaking and misusing of users’ personal health data. Especially, because of financial and operational benefits, the health service provider tends to outsource the huge number of health data and health query service to the cloud, who is able

to execute high-performance computing. As a result, protecting the users' privacy in health query has attracted considerable interest. On the other hand, the health data and diagnosis model are intellectual properties of the healthcare center, and should be kept away from the cloud. Lying at the intersection of mobile computing and cloud computing, designing a privacy-preserving health query on outsourced cloud faces the challenges below:

- **Challenges on accurate health query on encrypted data.** In the outsourced health query service, the health service provider and the user are not willing to disclose the sensitive health data and physiological data, so both the health data and the physiological data are encrypted before being sent to the cloud. Health query on encrypted data is a big challenge. Some privacy-preserving techniques like differential privacy may distort the data, making it unsuitable for medical analysis, sometimes even fatal for users [14]. Therefore, the health query scheme on encrypted data should be accurate for medical analysis.
- **Challenges on efficiency of health query on encrypted data.** Various cryptographic techniques have been proposed to conduct health query and data analysis on encrypted data. For example, different homomorphic encryption techniques have been reported [58–61]. But the overhead of the computation becomes a stumbling block to make them to be popularized at mass industrial level [104]. To provide good user experiences, the health query scheme on encrypted data should be efficient.
- **Challenges on Security.** Although the health data and users' physiological data are encrypted, it is reasonable to assume that an attacker knows the plaintext of some encrypted data, or some other side information. Thus, it is important to require the scheme to be robust against known-plaintext attack. Moreover, to prevent unauthenticated attackers, only authenticated users are allowed to conduct health query.

In this chapter, aiming at solving the above challenges, we propose an efficient and privacy-preserving health query scheme over outsourced cloud, named HeOC, which allows authenticated users to query on the sensitive health data without disclosing his/her sensitive physiological data to the cloud. Specifically, we design a novel anomaly detection technique, which efficiently detects anomaly disease with high risk for further analysis. The main contributions of this chapter are as following:

- First, we propose the HeOC scheme to enable an efficient privacy-preserving health query over outsourced cloud. Particularly, in this scheme, the user detects the anomaly disease with the sensor collected data first to filter out the suspicious disease with its related physiological data. Then through a variant of OPRF protocol (see Section 2.2.5), the user can query the accurate disease level with the filtered result from the sensor anomaly detection technique. Experiments demonstrate that our HeOC scheme is efficient.
- Second, we propose a novel sensor anomaly detection technique for detecting high-risk disease, called SADS. In the SADS technique, the health service provider outsources an encrypted health tree for reference to the cloud. Authenticated users send encrypted physiological data to the cloud to detect high-risk disease without disclosing his/her sensitive personal health data.
- Third, we develop an android app and two python programs to evaluate the performance of the HeOC scheme in real environment. The results show that the HeOC scheme is efficient with low computation and communication overhead.

The remainder of this chapter is organized as follows. In Section 4.2, we introduce our system model, security requirement and design goal. In Section 4.3, we present our HeOC

scheme, followed by its security analysis and performance evaluation in Section 4.4 and Section 4.5 respectively. Finally, we draw our conclusions in Section 4.6.

4.2 Models and Design Goal

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

4.2.1 System Model

Privacy-preserving health query has been studied in two settings of: two-party health query and outsourced health query. In this chapter, we focus on the outsourced health query setting. As shown in Fig.5.2, the system consists of three kinds of entities: health service provider, authenticated users and the cloud.

- Health service provider has abundant of health data. It can be professional medical organizations, hospitals. To simplify the system model, we take hospital as the health data owner. The hospital trains a number of diagnosis models for multiple diseases, and utilizes these diagnosis model to offer health query service. Because of the financial and operational benefits of data outsourcing, the hospital offers the health query via the cloud. However, the hospital is not willing to disclose the valuable diagnosis models to the cloud. Therefore, the hospital encrypts the diagnosis model and the health data, then outsources the encrypted data to the cloud.
- The users get authenticated by the hospital. Each authenticated user is equipped with a wearable BSN, which is comprised of a number of wearable sensor nodes wirelessly capturing and collaboratively processing the physiological data. A smart

Figure 4.2: System model of outsourced health query

phone is used to preprocess the collected data. To prevent user's privacy from leaking to the cloud, the user obtains an access control key in advance, utilizes cryptographic and privacy-preserving techniques to encrypt the query and sends it to the cloud.

- The cloud gets the encrypted query from users, utilizes the encrypted diagnosis model from the hospital to conduct disease diagnosis with cryptographic and privacy-preserving techniques. Then the cloud returns the encrypted diagnosis result to the users.

4.2.2 Security Requirement

In the context of outsourced system model, the cloud is considered as semi-honest, which means that the cloud would strictly execute the protocol to guarantee the correctness of the outsourced diagnosis service, but it also has financial incentives to recover the hospital's

valuable diagnosis model and user's privacy information. Moreover, the cloud may be compromised by hackers. Therefore, to guarantee the privacy of the user's physiological data and the confidentiality of hospital's diagnosis model, the following security requirements should be satisfied:

- *Privacy.* In the HeOC scheme, each user conducts health queries with sensor collected physiological and personal information like age, gender, medical history. The cloud should not know the user's physiological data by observing the ciphertext of the query data, which means that the HeOC scheme should be secure under ciphertext-only attack. Moreover, if attackers know the plaintexts of some encrypted user query data, the attackers should not be able to reveal the user's physiological data corresponding to other ciphertexts. In other words, the HeOC should be secure under known-plaintext attack too.
- *Confidentiality.* The hospital's diagnosis models should be prevented from being disclosed to the cloud, since these models are the intellectual properties belonging to the hospital. Same as above, the cloud should not be able to recover the diagnosis model with the encrypted medical data outsourced from the hospital, which means that the diagnosis models should be able to be secure under ciphertext-only attack. Furthermore, even when the attacker knows the plaintext of a particular encrypted disease diagnosis model, it should not be able to obtain the plaintext of the other encrypted diagnosis model, which means the HeOC scheme should be secure under known-plaintext attack.
- *Authentication.* Only authenticated users are allowed to send health query to the cloud and get the diagnosis result. If an illegal user forges a health query, this hacking operation should be detected immediately.

Besides the above security requirements, other attacks such as differential privacy attack, access-pattern attack may be possible. In this chapter, we only consider the security requirements above. We will leave the other attacks for future study.

4.2.3 Design Goal

Based on the system model mentioned above, our design goal is to develop an efficient and privacy-preserving health query over outsourced cloud. Specially, the following three objectives should be achieved:

- *Efficiency.* Considering the real-time requirements of online health query service and the diversity of the users, the proposed scheme should not consume too much computation and communication. Specifically, some time-consuming operations such as exponential operation should not be conducted to make the scheme acceptable.
- *Security.* The proposed scheme should be resilient to ciphertext-only attacks and known-plaintext attacks. The privacy of user's physiological data and confidentiality of the diagnosis model should be achieved.
- *Accuracy.* The accuracy of the health query should be guaranteed. The privacy enhancing techniques mentioned above should not be able to compromise the accuracy of the diagnosis service, because the inaccuracy may incur serious consequence for users. Therefore, the proposed scheme should achieve high accuracy.

4.3 Proposed HeOC Scheme

In this section, we propose a novel privacy-preserving disease query scheme based on a sensor anomaly detection technique, pseudorandom function PRFs (see Section 2.2.5) and

a symmetric encryption technique.

Heart beat	< 60	>=60,<80	>=80,<100	>=100,<120	>=120
81	0	0	1	0	0
59	1	0	0	0	0
88	0	0	1	0	0
101	0	0	0	1	0
125	0	0	0	0	1
68	0	1	0	0	0
130	0	0	0	0	1

Figure 4.3: Data discretization to turn numerical data into categories.

4.3.1 Overview

The proposed HeOC scheme allows detecting sensor anomaly and determining the disease level in a privacy-preserving way. In HeOC, each user is equipped with a wearable BSN, which is comprised of a number of wearable sensor nodes wirelessly capturing and collaboratively processing the user's physiological signals. These raw data are transmitted to the user's smart phone for further processing. The user's personal information includes age, gender, medical history, sensor data, etc. Before introducing the proposed scheme, we need to give an introduction about discretization of these user's personal information. Discretization is a process of transferring continuous data into discrete counterparts. We use *physiological data* to refer the standardized personal information. Let $\mathbb{X} = \{x_1 \dots x_n\}$ be a set of physiological data. Each element $x \in \mathbb{X}$ is standardized from the original personal information. Shown in Fig.4.3, for a heart rate from 80-100, x is 4 (00100); from

4.3. PROPOSED HEOC SCHEME

60-80, x is 8 (01000); below 60, x is 16 (10000); from 100-120, x is 2 (00010); above 120, x is 1 (00001). These preprocessing makes the physiological data suitable for numerical evaluation and allows to standardize the input of the algorithm, hence simplifying the implementation of the health query scheme.

Figure 4.4: Query flow in HeOC

The proposed HeOC scheme is comprised of a sensor anomaly detection technique (SADS) and a disease level query technique. Fig.4.4 depicts the query flow. The SADS technique allows the authenticated user to detect sensor data anomaly and determinate the related disease with high risk in a privacy-preserving way. Then, the user collects the physiological data related to the high risk disease, queries the specific disease level with the

help of the cloud in a privacy-preserving way. For the readers' convenience, we summarize the important notations to be used in Table 4.1.

Table 4.1: Notations frequently used in HeOC

Notation	Description
λ	security parameter
$Enc()$	public encryption algorithm generated by the hospital
$Enc_c()$	public encryption algorithm generated by the cloud
$\{x_a \dots x_b\} \in \mathbb{X}$	physiological data
$\{s_a \dots s_b\} \in \mathbb{S}$	sensors in user side
$hdmt$	healthy D-Merkle tree build
lnd	leaf-node dictionary
DL_{ij}	level j of disease i
$hdmt - c$	healthy D-Merkle tree outsourced to the cloud
$hash()$	hash function for building node in D-Merkle tree
$dmt - u$	D-Merkle tree build by the user
pf	pseudorandom function
k_1	pseudorandom function key
k_2	pseudorandom function key
k_3	user access symmetric key
$DLDic$	disease level dictionary

4.3.2 Proposed D-Merkle Tree

Before describing HeOC scheme, we give an introduction on the D-Merkle tree, which serves as a building block of the HeOC scheme. The proposed D-Merkle tree is a variant of Merkle tree. As shown in Fig.4.5, D-Merkle tree has the following properties: 1) each non-leaf node is labeled with the cryptographic hash value of the labels of its child nodes; 2) each leaf node represents a disease, with points to the related physiological data $\{x_a \dots x_b\} \in \mathbb{X}$, which are collected from the sensors $\{s_a \dots s_b\} \in \mathbb{S}$; 3) each leaf node has an accompany list, e.g. in Fig.4.5, the accompany list of the leaf node D_1 is $[D_2, n_2]$. Given one leaf node and its accompany list, the root of the D-Merkle tree can be easily recovered. In the context of the user-cloud-hospital disease level detection scheme, the hospital generates

Figure 4.5: An example of Disease Merkle Tree with physiological data $x_1, \dots, x_{10}$, which are collected from sensors $s_1, \dots, s_{10}$.

the D-Merkle tree, and outsources it to the authenticated users and the cloud. Specifically, for confidentiality issue, the D-Merkle tree outsourced to the cloud is preprocessed by the hospital shown as Fig.4.6. In the example of D-Merkle tree in Fig.4.5, if there were only three diseases D_1, D_2, D_3 , the D_3 node would be the sibling node of n_1 .

Moreover, the hospital builds a leaf dictionary, each (key,value) pair represents (nodeId, leafNode`Pointer). With this leaf dictionary, the cost to find a disease leaf node is $O(1)$. In the proposed HeOC scheme, it is high frequency to find a disease leaf node. Thus, this data structure will make the scheme efficient.

Figure 4.6: Data structures of nodes for D-Merkle tree outsourced to cloud

4.3.3 System Setup

First of all, the hospital sets up a public key encryption algorithm which is IND-CCA2 secure (see Section 2.2.6) [105], and publishes the public key and the encryption algorithm $Enc()$. The cloud also sets up a public key encryption algorithm which is IND-CCA2 secure, and publishes the public key and the encryption algorithm $Enc_c()$. Then, the hospital builds a diagnosis table according to multi-level diagnosis model for multi-disease. In each row of the diagnosis table, there is a disease level DL_{ij} and corresponding physiological data $\{x_a \dots x_b\}$. Given practical consideration, the physiological data $\{x_a \dots x_b\} \in \mathbb{X}$ are standardized as mentioned in section 4.3.1, and each value $x \in \mathbb{X}$, $x < MaxLevel$, $MaxLevel$ is the maximum number of levels for disease. To simplify the notation, in each row of the diagnosis table, for level j of disease i , we note it DL_{ij} , which has related physiological data $\{x_a \dots x_b\} \in \mathbb{X}$.

Hospital Builds The Healthy D-Merkle Tree

In the diagnosis table, the hospital picks the healthy level of all the diseases, and builds a healthy D-Merkle tree $hdmt$ and a leaf-node dictionary lnd . For each leaf node, a node id is

used as the pseudonym in the sensor anomaly detection technique later. Given d diseases, the cost for searching the $hdmt$ is $O(\log d)$. With the dictionary lnd , the cost to find the leaf node for disease level DL_{ij} is $O(1)$.

For confidentiality, before the D-Merkle tree $hdmt$ and the leaf-node dictionary lnd are outsourced to the cloud, the tree nodes are preprocessed as mentioned in section 4.3.2. As a result, the D-Merkle tree in the cloud does not contain any disease name and physiological pointer array in the leaf nodes. We denote this tree $hdmt - c$. To fulfill the check algorithm later, the cloud uses the node id to distinguish the leaf nodes. The procedure to build the $hdmt - c$ from the $hdmt$ is shown in Fig.4.6. In the procedure, k_3 is a symmetric key, used for user access control.

User Registration

The user registration algorithm is shown in Algorithm 1. When a user registers for privacy-preserving health query, he or she first collects his/her sensor information, then sends the sensor list to the hospital. The hospital selects the diseases that can be detected according to the user's sensor list, and retrieves the disease ids related to the selected diseases from the leaf node of the $hdmt$. Then the hospital generates a disease id list with these disease ids , and returns the list disease list back to the user. For instance, if the hospital owns the $hdmt$ as Fig.4.5, and receives the sensor list (s_3, s_4, s_8, s_9) , the hospital will returns the disease id list (D_1, D_4) back to the user.

Additionally, the user learns the relationship between each disease id with its corresponding sensors from the hospital. The user also gets the hash value of the root of the $hdmt$, $hash_{root}$ for verification later in the sensor anomaly detection algorithm. Moreover, the user gets the symmetric key k_3 and the hash function $hash()$ from the hospital through

a secure channel.

Algorithm 1 Diseases_from_Sensors()

```

1:  $\setminus\setminus (S_1, S_2, \dots, S_k)$  sensors
2:  $\setminus\setminus Ls$  bit length of a sensor  $S_i$ 
3:  $\setminus\setminus k$  size of the sensors
4: User does:
5:  $r = rand()$ 
6: for  $i = 1$  to  $k$  do
7:    $Enc(S_i|r)$ 
8: end for
9: sends  $(Enc(S_1|r), Enc(S_2|r), \dots, Enc(S_k|r))$  to the hospital
10: 

---


11: Hospital does:
12: for  $i = 1$  to  $k$  do
13:    $S_i|r = Dec(Enc(S_i|r))$ 
14:    $S_i$  = first  $Ls$  bits of the  $S_i|r$ 
15:    $r$  = rest bits of the  $S_i|r$ 
16: end for
17: find out  $(D_1, D_2, \dots, D_n)$  according to the  $(S_1, S_2, \dots, S_k)$ 
18: sends  $(D_1 \oplus r, D_2 \oplus r, \dots, D_n \oplus r)$  to the user
19: 

---


20: User does:
21: for  $i = 1$  to  $n$  do
22:    $D_i = D_i \oplus r \oplus r$ 
23: end for
24: User gets  $(D_1, D_2, \dots, D_n)$ 
  
```

Hospital Builds An Encrypted Diagnosis Database

To fulfill the privacy-preserving disease level query, the hospital builds an encrypted database based on the diagnosis table. The hospital takes λ as the security parameter. We adopt one of the latest searchable symmetric encryption construction [106]. The function pf is a variable-input-length pseudorandom function (PRF) (see Section 2.2.5) on input key $K \in \{0, 1\}^\lambda$, $x \in \{0, 1\}^*$, output a string in $\{0, 1\}^\lambda$. The Hospital picks two PRF keys $k_1, k_2 \in K$ for PRF pf , and builds the encrypted disease level index. For each disease

level DL_{ij} , the hospital encrypts it as below:

$$\begin{cases} T_1 = pf(k_1, x_a || \dots || x_b) \\ T_2 = pf(k_2, x_a || \dots || x_b) \oplus DL_{ij} \end{cases}$$

In the equation above, for each disease level DL_{ij} , all the related physiological data $(x_a, \dots, x_b) \in \mathbb{X}$ are concatenated to be the input $x_a || \dots || x_b$ of the PRF pf . Each row of the encrypted database is hashed by T_1 as key, T_2 as value, which are added in the dictionary $DLDic$. Then the hospital outsources the dictionary $DLDic$ to the cloud.

After the setup, the cloud gets the preprocessed healthy D-Merkle tree $hdmt - c$, encrypted disease level dictionary $DLDic$. Each registered user gets the disease id list according to his/her sensor information, the hash value of the root of the $hdmt$, $hash_{root}$, the access key k_3 , the hash function $hash()$, the security parameter λ , two PRF keys $k_1, k_2 \in K$ and the PRF pf .

4.3.4 Sensor Anomaly Detection Technique

In this section, we present a novel sensor anomaly detection technique named SADS, which will serve as the basis of our HeOC scheme for privacy-preserving health query over outsourced cloud.

Monitor Algorithm

The BSN placed in different location on the human body wirelessly captures and collaboratively processes the user's physiological data, synthesizes responses and forwards the data to the user's smartphone. The user inputs his/her gender, age, medical history, with the collected sensor data. These data are standardized to be physiological data $\{x_1 \dots x_n\}$ in the smartphone. According to the relationship between the disease ids and sensors got

from the hospital, the user calculates the hash value for each disease id . For a disease id id_i and the physiological data $\{x_1, x_2, \dots, x_k\}$ collected by its associated sensors $\{s_1, s_2, \dots, s_k\}$, the user calculates $hash_i = hash(x_1|x_2|\dots|x_k|k_3)$.

In the monitor algorithm, for the disease id vector $Id = (id_1, id_2, \dots, id_n)$ got from the hospital, the user calculates the hash vector

$$Hash = (hash_1, hash_2, \dots, hash_n)$$

Verification and Check Algorithm

When a user wants to find an anomaly disease with high risk, he or she needs to download the disease information from the cloud to conduct the comparison. The downloaded information can be verified with the hash value of the root of the $hdmt$, $hash_{root}$. Different users may have different sensors, so the disease information downloaded from the cloud varies according to the sensors the user owns.

The algorithm to download the disease information from the cloud is shown in the Algorithm 2. Considering the practicality of the different sensors the users wear, each authenticated user downloads the leaf nodes ($LeafList$) of the $hdmt - c$ related to the user's sensors. Additionally, the user downloads the accompany list for each leaf node to verify these leaf node. In the algorithm, $AccompList$ stores the accompany nodes for each leaf node. At the end of this algorithm, the cloud returns the $LeafList$ and $AccompListList$ back to the user.

After the user downloads the leaf nodes and the correspondingly accompany list, the user compares and verifies the information in the user-side, shown as the Algorithm 3. In the algorithm, the user first verifies all the downloaded leaf nodes by computing the

4.3. PROPOSED HEOC SCHEME

Algorithm 2 Search_Diseases(*DiseaseList*, *hdmt* – *c*)

```
1: \\ DiseaseList,  $\{D_1, D_2, \dots, D_n\}$  disease list owned by the user.
2: \\ Output: LeafList, the hash value of the leaf nodes searched.
3: \\ Output: AccompListList, list of the  $n$  accompany list for DiseaseList =  $\{D_1, D_2, \dots, D_n\}$ 
4: for  $i = 1$  to  $n$  do
5:   List AccompList
6:    $node = hdmt - c.leaf(D_i)$ 
7:   LeafList.add(node)
8:   AccompList.add(node.sibling)
9:   while  $node \neq root$  do
10:     $node = node.parent$ 
11:    AccompList.add(node.sibling)
12:   end while
13:   AccompListList.add(AccompList)
14: end for
15: Cloud sends AccompListList and LeafList to the user
```

hash value of each leaf node and the nodes of its accompany list. If all the downloaded information are correct, the computed hash value should be same as the hash value of the root of the $hdmt - c$, $hash_{root}$. Then, the user compares his hash value $hash_i \in Hash = (hash_1, hash_2, \dots, hash_n)$ obtained from the Monitor Algorithm with the counterpart hash value in the downloaded leaf node. If these two values are different, it means the user has high risk for this disease. Thus, the disease information D_i will be added into the *AnomalyDiseaseList* for further detection.

After the verification and check algorithm, the user verifies the downloaded leaf nodes and filters out the diseases with high risk, which are stored in the *AnomalyDiseaseList*.

4.3.5 Privacy-preserving Disease Level Query

The authenticated user gets the key k_1, k_2 from the hospital through a secure channel. With the SADS technique, the user finds out the anomaly leaf nodes, each of which represents one disease. To simplify the explanation of the disease level query technique, we consider one leaf node and one corresponding disease. The query technique is same for the other

Algorithm 3 Verify_and_Check()

```

1: \ Input: AccompListList, hashroot, LeafList
2: \ Input: hashi ∈ Hash = (hash1, hash2, ..., hashn)
3: \ Output: AnomalyDiseaseList, anomaly leaf nodes
4: for i = 1 to n do
5:   AccompList = AccompListList.get(i)
6:   leafnode = LeafList.get(i)
7:   for j = 1 to AccompList.size() do
8:     hashr = leafnode.hash
9:     hashl = AccompList.get(j).hash
10:    hashr = hash(hashl ⊕ hashr)
11:   end for
12:   if hashr ≠ hashroot then
13:     alert( verification error)
14:   end if
15:   if leafnode.hash ≠ hashi then
16:     AnomalyDiseaseList.add(Di)
17:   end if
18: end for

```

anomaly leaf nodes.

For an anomaly leaf node and the corresponding disease, the user gets the related ($x_a, \dots, x_b$) from the *physiological_point_array* in the leaf node. With pseudorandom function pf , the user calculates two values as follows:

$$\begin{cases} t_1 = pf(k_1, x_a || \dots || x_b) \\ t_2 = pf(k_2, x_a || \dots || x_b) \end{cases}$$

generates a random value r and encrypts t_1 with the cloud's public key encryption algorithm, $Enc_c(t_1|r)$. Then the user sends $Enc_c(t_1|r)$ to the cloud.

The cloud decrypts the $Enc_c(t_1|r)$ and gets t_1 and r as below:

$$t_1|r = Dec_c(Enc_c(t_1|r))$$

Because t_1 is the length of λ , t_1 is the left λ bits of the $t_1|r$, and the remaining bits of the $t_1|r$ is the random number r . The cloud uses t_1 as the dictionary key to search the value in the disease level dictionary $DLDic$, and gets the result ($\perp$ or T_2). $\perp$ means no matched value was found. Then, the cloud calculates $T_{2r} = T_2 \oplus r$, and returns ($\perp$ or T_{2r}) back to the user.

If the query result is not $\perp$, the user calculates the disease level as:

$$\begin{aligned}
 DL &= r \oplus t_2 \oplus T_{2r} \\
 &= r \oplus t_2 \oplus (T_2 \oplus r) \\
 &= t_2 \oplus pf(k_2, x_a || \dots || x_b) \oplus DL_{ij} \\
 &= pf(k_2, x_a || \dots || x_b) \oplus pf(k_2, x_a || \dots || x_b) \oplus DL_{ij} \\
 &= DL_{ij}
 \end{aligned}$$

where DL_{ij} is the level j of disease i .

4.3.6 Correctness Analysis

First, we prove that when the user sends his/her sensor information, he/she can get the correct related disease information from the hospital through the HeOC scheme. Each sensor $S_i \in \{S_1, S_2, \dots, S_k\}$ is encrypted by the user as $Enc(S_i|r)$. r is a random number generated by the user. When receiving $Enc(S_i|r)$, the hospital can decrypt the information and get the $S_i|r$, because the hospital has the private key of the public key encryption technique. Moreover, because each $S_i \in \{S_1, S_2, \dots, S_k\}$ is L_s bits in length, the hospital can easily get the S_i and the random number r . After finding out the diseases $\{D_1, D_2, \dots, D_n\}$

according to the sensors $S_1, S_2, \dots, S_k$, the hospital returns the $\{D_1 \oplus r, D_2 \oplus r, \dots, D_n \oplus r\}$ back to the user. Because the random number r is generated by the user, the user can easily recover the correct $\{D_1, D_2, \dots, D_n\}$.

Second, we prove that the user can correctly verify the leaf nodes downloaded from the cloud in the SADS technique. In the setup, the hospital publishes the root of the *hdmt*, $hash_{root}$ as the signature of the *hdmt*. When a user wants to detect the anomaly diseases, he/she downloads all the leaf nodes and the related accompany list according to his/her disease list. Shown as the line 6 to line 11 in the Algorithm 3, the user can calculate a new hash value $hash_r$. If the $hash_r$ is same as the public signature $hash_{root}$, the related leaf node is proved to be correct.

Third, we prove that the user can correctly get the disease level DL_{ij} by querying with $t_1 = pf(k_1, x_a || \dots || x_b)$. In the system setup, the hospital encrypts the disease diagnosis table and builds the disease level dictionary *DLDic*. Each row of the diagnosis table contains information for level j of disease i DL_{ij} with the related physiological data $(x_a, \dots, x_b) \in \mathbb{X}$. In the dictionary *DLDic*, the key is $pf(k_1, x_a || \dots || x_b)$, and the value is $pf(k_2, x_a || \dots || x_b) \oplus DL_{ij}$. Therefore, in the disease level query technique, the user concatenates the $(x_a, \dots, x_b)$ and calculates the key $pf(k_1, x_a || \dots || x_b)$, which is same as the key build by the hospital. Therefore, the hospital will definitely return the responding value $pf(k_2, x_a || \dots || x_b) \oplus DL_{ij} \oplus r$. Because the random number r is generated by the user, when receiving the query result, the user removes the $pf(k_2, x_a || \dots || x_b) \oplus r$ with the xor operation with t_2 and r , and gets the DL_{ij} , which is the correct level j of disease i .

4.4 Security Analysis

In this section, we analyze the security properties of the proposed HeOC scheme. Specifically, our analysis will focus on how the proposed HeOC scheme can achieve the user's personal data privacy and the hospital's diagnosis model confidentiality.

The user's personal privacy involves the user's physiological data, medical history, age, gender, etc. On the other hand, hospital diagnosis model confidentiality involves the disease level with its related medical data in the encrypted D-Merkle tree and disease level query technique. The security of the HeOC scheme relies on the underlying encryption, hash function, pseudorandom function and the SADS technique. The pseudorandom function prevents the cloud from learning the disease diagnosis model database while the SADS technique and the disease query technique protect the user's private personal information and hospital diagnosis model from the cloud. Thus, we focus on the diagnosis of user's privacy and disease model in three stages: system setup, sensor anomaly detection and disease level query.

Security of the system setup. In the system setup, a healthy D-Merkle tree is build by the hospital with the physiological data in the healthy status. In the non-leaf nodes of the outsourced healthy D-Merkle tree $hdmt - c$, all the names are removed as shown in Figure 4.6. And in the leaf nodes of $hdmt - c$, all the names and physiological pointer are deleted. Therefore, only the hash value and id are kept to identify the nodes. The attacker can not decipher the diseases related to the leaf nodes and the corresponding physiological data $(x_a, ..., x_b) \in \mathbb{X}$.

On the other hand, the IND-CCA2 secure public encryption technique $Enc()$ prevents the sensor information from being hacked. Additionally, when the hospital returns the disease list, the random value r generated by the user hides the disease list against attackers.

Security of the sensor anomaly detection. In the sensor anomaly detection technique, the user downloads all the leaf nodes and the related accompany list according to his/her disease list. In the disease list, only the disease ids are used as the pseudonyms, so that the disease list will not leak any sensitive information of the user. Even when an attacker knows one disease with its corresponding disease id, the attacker can not recover the other disease information of the user. Moreover, the user verifies and checks high risk diseases on his/her own. The attacker can not recover the user's sensitive physiological data.

Security of the disease level query. As discussed above, user's personal information and hospital diagnosis model in sensor anomaly detection technique are secure. It is also hard to attack these privacy in the disease level query technique, because the database are encrypted with $pf(k_1, x_a || \dots || x_b)$, $pf(k_2, x_a || \dots || x_b) \oplus DL_{ij}$. We follow the constructions in [106], which is able to achieve efficient constructions for pseudorandom function and encryption scheme to guarantee chosen-keyword attacks (CKA2). Same as above, the IND-CCA2 secure public key encryption technique Enc_c generated by the cloud prevents the $t_1|r$ from being recovered by the hackers. The random number r generated by the user keeps the disease level DL_{ij} stored in the $T_{2r} = T_2 \oplus r$ away from being recovered by the hackers.

4.5 Performance Evaluation

In this section, we evaluate the performance of the proposed HeOC scheme in terms of computational cost and communication cost.

4.5.1 Implementation and Experimental Settings

We have implemented the HeOC scheme in Python and Java. We tested HeOC's performance in a testbed of Alibaba Elastic compute service, a Mac OSX laptop and one Android phone. Those machines play the roles of cloud, hospital and mobile user. The hardware and software of these machines are shown in Table 4.2, while the parameter settings are shown in Table 4.3. Specifically, we adopt the Goldreich-Goldwasser-Micali [107] and SHA256 to construct PRFs.

Table 4.2: Experimental setting

Role	Machine	Hardware & Software
Cloud	Alibaba ECS	Instance ecs.xn4.small, CentOS 7.2 64-bit and Python
Health provider	Mac laptop	CPU:2.9 GHz Intel Core i5, memory: 8GB
User	MEIZU phone	CPU: Exynos 7872 3 GB ram; Android 6.0.1

Table 4.3: Parameter setting

Parameter	Setting
λ	256
length of <i>hash()</i> output	256 bits
hash for D-MerkleTree	sha256

4.5.2 Hospital's computational cost

In the system setup, the hospital builds a healthy D-Merkle tree *hdmt* and encrypts the diagnosis table. According to a survey of World Health Organization (WHO), the WHO distinguishes 12,420 disease categories in ICD-10 (2010) [108]. Thus, we test the time for SADS and disease level query with disease level number of 10^5 , $2*10^5$, $3*10^5$, $4*10^5$, $5*10^5$, shown in Fig.4.7. Even for $5 * 10^5$ disease, it takes less than 2.52 seconds to build the *hdmt*

for the hospital. Considering that system setup is conducted only once, the setup time is acceptable for the hospital.

Figure 4.7: Hospital latency for building the *hdmt*.

4.5.3 Computational cost at user side

The user has three tasks: 1) The user finds the disease list according to the user's sensors in user registration. 2) The user downloads the leaf nodes and related accompany list according to the disease list, then verifies and checks the downloaded information to find the disease with high risk in the SADS technique. 3) The user uses *PRF* to detect the accurate disease level DL_{ij} for each disease with high risk in the disease level query technique.

To test the user registration latency, we let the android phone generates 100, 300, 500, 700, 900 sensors. The test result is shown in Fig.4.8. It takes less than 6.8 seconds even with 900 sensors to initialize the setup for the user. The user registration is conducted only once for the user, so the computational cost is acceptable.

4.5. PERFORMANCE EVALUATION

To clearly test the latency in the SADS for the user, the most prominent time is the latency for verification. The verification latency for each disease is strongly related to the tree level of the *hdmt* stored in the cloud. Thus, we test the verification latency with the 10^2 , 10^3 , 10^4 , 10^5 , 10^6 diseases in the cloud. The result is shown in Fig.4.9. The test result demonstrates that the verification latency climbs gradually with the tree level of the *hdmt*. Even with 10^6 diseases, the tree level of the *hdmt* is 22, the verification latency for one disease is only 3 milliseconds. Therefore, one user with hundreds of diseases to verify only needs about 1 seconds to verify and check all the information to find the diseases with high risk. It is very efficient. Given that the user has n diseases to test and the tree level of the *hdmt* is L , the time complexity of the user side is $O(nL)$.

Two PRF pf encryptions are required for the user to query the disease level DL_{ij} . From the test result, we know that it takes less than 27 milliseconds to encrypt an anomaly leaf with pf , which is negligible for a user with smartphone.

Figure 4.8: User latency for downloading diseases.

Figure 4.9: User verification latency and related *hdmt* tree level.

4.5.4 Computational cost at cloud side

In the SADS technique, the user downloads all the leaf nodes and the related accompany list in the *hdmt* from the cloud. The cloud latency in the SADS is strongly related to the tree level and the queried disease number. To clearly test the cloud latency, we test the cloud latency for one queried disease while the *hdmt* has 10^2 , 10^3 , 10^4 , 10^5 , 10^6 diseases. The result is shown in Fig.4.10. Assuming a user has 1000 diseases to query in the SADS technique, it shows that even with 10^6 diseases in the *hdmt*, the computational cost to search the anomaly leaf node at the cloud side is acceptable.

To find the high risk leaf node, the cloud searches the value with the key $t_1 = pf(k_1, x_a || \dots || x_b)$. The computational cost to find a value in the dictionary *DLDic* is $O(1)$, which is negligible for the cloud.

Figure 4.10: Cloud latency for the user to query one disease

4.5.5 Communication Cost

In the system setup, the hospital outsources the *hdmt* and the encrypted disease level database to the cloud. Each tree node of *hdmt* includes one 32 bits node *id* and one 256 bits hash value. Each row of the encrypted disease level has one 256 bits key and one 256 bits value. Considering we have n diseases and total L disease levels, the total communication cost for hospital to outsource the *hdmt* and the encrypted disease level is about $(2n - \frac{n}{2^{\log n}}) * 288 + 512L \approx (2n - 1) * 288 + 512L$ bits under the system settings in Table 4.3. It is conducted only one time. So it is acceptable for the hospital.

In the user registration, a user sends all his/her sensor information to query the diseases that can be detected. Assuming the user has k sensors and the query result is n diseases that will be detected. Let each sensor and each disease be 32 bits in length, it takes $32k$ bits to send the sensors and $32n$ bits to receive the query result.

In the SADS technique, a user sends the disease list to the cloud to download the leaf

nodes and the related accompany list in the *hdmt*. Let the *hdmt* in the cloud has D diseases and the user has n diseases to detect, it takes $32 * n * (\log D + 1)$ bits to download all the leaf nodes and the accompany list. In reality, the disease number n that the user has will be at most 1000. The disease number D in the *hdmt* will be around 10^6 . Thus, the communication cost is acceptable for the user to download the leaf nodes and the related accompany list.

4.6 Conclusions

In this chapter, we have proposed an efficient and privacy-preserving health query over outsourced cloud, which preserves the user's personal privacy and achieves the confidentiality of the health service provider's diagnosis model. To realize HeOC, we have designed a sensor anomaly detection technique to find the high risk diseases, then with oblivious pseudorandom function protocol, the user queries the health diagnosis result accurately. In addition, performance evaluation demonstrates its efficiency in computation and communication overhead. According to our tests, a user with hundreds of diseases to verify only needs around 1 second to verify and check all the information to find the disease with high risk. It is very efficient. At the cloud side, even with 10^6 diseases, the computational cost to search the anomaly leaf node is around 0.06 ms.

Chapter 5

Privacy-Preserving Priority Classification on Patient Health Data in Remote e-Healthcare System

In the last chapter, we have discussed the three-party system model, in which, the healthcare center outsources the computation tasks to the cloud server, and the users conduct healthcare query from the cloud. Both of the proposed schemes in the Chapter 3 and the Chapter 4 are interactive schemes, which need the service provider or the cloud to achieve the privacy-preserving tasks with help from the users. In this chapter, we will discuss a non-interactive medical-packet-relay model in the remote e-healthcare system, which is comprised of users, gateways and healthcare center. In this following non-interactive system, each gateway achieves the privacy-preserving tasks by itself.

5.1 Introduction

A variety of wireless body area network (WBAN) schemes and applications have been proposed [109–112] in recent years, including energy-efficient medium access protocol for WBAN using the listen-before-transmit manner [110], data forwarding framework between biosensors and the gateway considering the presence of body shadowing [111], prioritized adaptive resource allocation algorithm for WBAN based on patients' medical situation [112]. Considering the limited resource of the sensors, the collected data streams usually can not be transmitted directly to the healthcare center. As shown in Fig. 5.1, the sensors in each user's wearable health system periodically collect the users' physiological data, send these raw data to the his/her smartphone for preprocessing. The smartphone assembles a medical packet containing the user's preprocessed physiological data, and sends it to a WBAN-gateway nearby. The medical packets from different users will be aggregated in WBAN-gateways. Then the WBAN-gateways relay all the medical packets to the remote healthcare center.

Such WBAN remote healthcare system typically require users to submit sensitive physiological data, personal information like age, name, gender, medical history, which seriously raises concerns about leaking and misuse of users' sensitive privacy data. On the other hand, the healthcare center is not willing to reveal the disease models to the users or the WBAN-gateways. Some attackers may crack the users' smartphones or the WBAN-gateways, and steal the sensitive users' personal information and the healthcare center's intellectual properties. Therefore, a variety of privacy-preserving schemes have been proposed in remote e-healthcare system [69–72]. However, the privacy-preserving healthcare schemes based on the encrypted data have some issues like accuracy and efficiency to be solved. We outline the challenges for privacy-preserving remote e-healthcare system would

Figure 5.1: Wearable health monitor system

face as below:

- Challenges on security and privacy. Similar to the security and privacy issues discussed in the chapter above, all the users' physiological data, personal information and the healthcare center's disease models need to be encrypted. An attacker should not recover the sensitive plaintext by observing the ciphertext, i.e., secure under ciphertext-only attack. Moreover, it is reasonable to assume in some scenarios, the attacker knows some users' information or some disease models. Even in this context, the attacker can not recover other plaintext of the corresponding encrypted data. In other words, the system should be secure under know-plaintext attack.
- Challenges on accuracy. Similar to the accuracy issue discussed in the chapter above, to achieve the security requirements of the remote e-healthcare system, some privacy-preserving schemes based on the encrypted data need to standardize the users' personal information and healthcare center's disease model first. The standardization

techniques may compromise the computational accuracy. What's more, some randomization techniques like the differential privacy [13] add some random values to the computational results, which may cause medical disaster in some scenarios [14]. Therefore, the privacy-preserving remote e-healthcare system should be accurate for medical analysis.

- Challenges on efficiency. Different from the efficiency issues discussed in the chapter above, the system model in this chapter is non-interactive, which are always associated with time-consuming techniques. Most of the privacy-preserving schemes based on the encrypted data are involved with large computational overhead. Recent studies propose some privacy-preserving schemes with multi parties [15] is efficient, however, it can not be utilized in the non-interactive scenario. Thus, the privacy-preserving remote e-healthcare system needs to solve the efficiency issues.

In this chapter, aiming at solving the above challenges, we propose an efficient and privacy-preserving priority classification (PPC) on patient health data in remote e-healthcare system, which allows authenticated users to periodically send medical packets to the healthcare center through WBAN-gateways. The WBAN-gateways relay these aggregated medical packets in a non-interactive privacy-preserving way based on the packets' priorities (criticalities). The main contributions of this chapter are as following:

- First, we propose the PPC scheme, an efficient privacy-preserving non-interactive priority classification scheme for users' medical packets in WBAN-gateways. Particularly, The WBAN-gateways derive the priorities of the medical packets and relay the packets in a priority heap.
- Second, we develop an android app and two java server programs to evaluate the

performance of the PPC scheme. The results show that the proposed PPC scheme is efficient in both computational cost and communication overhead. The security analysis also demonstrates that our proposed PPC scheme can preserve the privacy of the users' personal information and the confidentiality of the healthcare center's disease models.

The remainder of this chapter is organized as follows. In Section 5.2, we formalize our system model, security requirement and design goal. In Section 5.3, we present our PPC scheme, followed by its security analysis and performance evaluation in Section 5.4 and Section 5.5 respectively. Finally, we draw our conclusions in Section 5.6.

5.2 Models and Design Goal

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

5.2.1 System Model

Privacy-preserving remote e-healthcare system has been studied in both cloud-based outsourced setting and two-party communication setting between user and healthcare center. In our work, we focus on the communication between the users and the healthcare center through relay WBAN-gateways. The WBAN-gateways receive medical packets from different users, and relay these packets on basis of the criticalities of the medical packets and their waiting time. As shown in Fig. 5.2, the system mainly contains three entities: healthcare center, WBAN-users, and WBAN-gateways.

Figure 5.2: System model of medical packet classification and relay in remote e-healthcare system under consideration

- *Healthcare center*: The healthcare center is a medical organization, which is professional in offering healthcare service, and has abundant diagnosis models for different diseases. The healthcare center offers custom medical service for different users. In other words, it provides individual diagnosis service for different diseases for different users according to the users' equipped sensors. Moreover, it sets different criticality thresholds for different users with the same disease because of their different physical conditions. However, the disease models and the thresholds are the intellectual properties of the healthcare center. The healthcare center is not willing to disclose the valuable intellectual properties to the users and the WBAN-gateways. Thus, it encrypts all the data, and sends the encrypted disease models and the thresholds to the authenticated users.
- *WBAN-users*: The users who equipped with a list of body sensors and smartphones

register with the healthcare center, send their sensor list and personal information to the healthcare center and retrieve the encrypted disease models and thresholds. The sensors periodically collect a user's physiological information, and send these raw data to the user's smartphone for preprocessing and encryption. Then, the user sends the encrypted medical packets to a WBAN-gateway. The WBAN-gateway will relay the packets to the healthcare center.

- *WBAN-gateways*: The WBAN-gateways are responsible for relaying users' medical packets to the healthcare center. The users' packets are randomly aggregated at the WBAN-gateways. These medical packets' transmission to remote healthcare center rely on a priority-based classification scheme managed by the WBAN-gateways. To simplify the description of our proposed scheme, we consider only one WBAN-gateway in our proposed scheme.

5.2.2 Security Requirement

Different from the security requirements for multiparty system discussed in the last chapter, in this system, the WBAN-gateway is considered as semi-honest, which means the WBAN-gateway would strictly execute the protocol to guarantee the correctness of the medical packets relay task, but it also has financial incentives to recover the healthcare center's valuable disease model, the users' privacy information. Moreover, the WBAN-gateway may be compromised by hackers. Therefore, to guarantee the privacy of the users' physiological data and the confidentiality of the healthcare center's disease model, the following security requirements should be satisfied:

- *Privacy*. In our remote e-healthcare system, each user collects his/her physiological

information and personal information, and sends these information to the healthcare for remote healthcare monitoring. These sensitive personal data should be prevented from leaking to the WBAN-gateway. Specifically, the WBAN-gateway can not recover the users' physiological data by observing the encrypted medical packets, which means our system is secure under ciphertext-only attack. What's more, if an attacker knows the plaintext of some encrypted medical packets, the attacker can not reveal a user's physiological data corresponding to other encrypted medical packets. In other words, it is secure under known-plaintext attack.

- *Confidentiality.* The disease models and the thresholds for each user are the intellectual properties of the healthcare center. Same as above, the WBAN-gateway can not recover the disease models and the thresholds by observing the encrypted data. In our system, the encrypted disease models and thresholds would be sent to the authenticated users. The authenticated users are allowed to make use of the encrypted data to conduct the preprocessing for the privacy-preserving priority classification, but can not recover the disease models and the thresholds. Moreover, even when an attacker knows some plaintext of one encrypted disease model, it is unable for the attacker to reveal other disease models, thresholds and the users' personal data, which is secure under known-plaintext attack.
- *Authentication.* Only the users, who have registered in the healthcare center and gained the the encrypted parameters, can conduct the remote healthcare monitoring service. The DDoS attack is not considered in our chapter. We focus on how to achieve the privacy-preserving priority classification task.

Note that, other attacks such as differential privacy attack, access-pattern attack and DDoS attack could be possible in e-healthcare system. However, since this work focus on

the privacy-preserving medical packets classification, those attacks will not be discussed, and would be exploited in our future work.

5.2.3 Design Goal

Based on the system model mentioned above, our design goal is to develop an efficient and privacy-preserving priority classification in remote e-healthcare system. Specially, the following three objectives should be achieved:

- *Efficiency.* Considering the complexity for non-interactive privacy-preserving computation for the WBAN-gateway and the real-time requirements of the emergency healthcare remote monitoring system, the proposed scheme should be efficient in computation and communication. In our system, some medical packets come from users with very critical situation. Thus, the privacy-preserving priority calculation for the packets in the WBAN-gateway should be efficient.
- *Security.* The aforementioned security requirements should be satisfied. The healthcare center's disease models and thresholds should not be recovered by the users and WBAN-gateway. On the other hand, the users' physiological data and personal information should be prevented from the WBAN-gateway.
- *Accuracy.* Because of the criticality of the e-healthcare system, the accuracy of the priority classification scheme in the remote e-healthcare system should be guaranteed. In some privacy-preserving schemes, the accuracy is compromised. However, in this system related to users' emergency healthcare monitoring, the privacy-preserving priority classification scheme should achieve high accuracy.

5.3 Proposed Privacy-Preserving Priority Classification Scheme

In this section, we propose a non-interactive privacy-preserving PPC scheme, and its system setup, user registration, user data collection, WBAN-gateway packet classification and packet relaying steps.

5.3.1 Overview

The proposed PPC scheme allows the WBAN-gateway to manage users' medical packets based on their emergency levels in a privacy-preserving way.

In PPC, each user is equipped with a wearable WBAN, which is comprised of a number of wearable sensor nodes wirelessly capturing and collaboratively processing physiological signals. These raw data are transmitted to the user's smartphone for standardization, encryption before being relayed to the WBAN-gateway as shown in Fig. 5.2. Some raw data are very small decimal numbers. To simplify the implementation of the encryption technique, the decimal data need to be increased by a factor of 1000 times and rounded up. We let $\mathbb{X} = \{x_1 \dots x_n\}$ be the standardized physiological data in the PPC scheme. Considering the decimal data like the AC conductance for skin conductance [113], for instance, the original $0.65 \mu S$ needs to be increased to $x = 650$. The standardization keeps the physiological data positive. Moreover, it ensures the inner product of the user's physiological data and the disease model offered by the healthcare center positive, which is an important property for our privacy-preserving emergency relay scheme. On the other hand, the healthcare center owns the disease model $(a_1, a_2, \dots, a_k)$ and thresholds $(t_1, t_2, \dots, t_l)$. The disease risk score S for the user, can be computed by $S = \sum_{i=1}^k a_i \cdot x_i$ [20, 114, 115]. We can calculate a user's disease risk by comparing the disease risk score and the thresholds $(t_1, t_2, \dots, t_l)$.

Before describing the PPC scheme, we define the priority classification on patients'

5.3. PROPOSED PRIVACY-PRESERVING PRIORITY CLASSIFICATION SCHEME

health data in remote e-healthcare system without considering the users' privacy and healthcare center's disease model confidentiality. The system consists of three stages:

- Each authenticated user retrieves the disease model $(a_1, a_2, \dots, a_k)$ and thresholds $(t_1, t_2, \dots, t_l)$ from the healthcare center according to the user's sensor list.
- The user's sensors collect the physiological data and send these data to the user's smartphone. Then the user conducts the calculation $sum = \sum_{i=1}^k a_i \cdot x_i$. The user sends a medical packet containing the physiological data $(x_1, x_2, \dots, x_k)$, user information *userinfo*, the calculation result *sum* and the thresholds $(t_1, t_2, \dots, t_l)$ to a WBAN-gateway. The user can calculate the priority for the packet by himself/herself, but to clearly describe the same data flow over encrypted data in the PPC scheme, we let the WBAN-gateway calculate the priority of the medical packet.
- As shown in Fig. 5.3, after receiving the medical packet from the user, the WBAN-gateway picks two consecutive thresholds t_i, t_j , where $t_i < sum < t_j$, and assigns j to this medical packet's priority. If *sum* is larger than the largest threshold $sum > t_l$, the WBAN-gateway assigns $l + 1$ to the packet's priority. Finally, the WBAN-gateway relays all the aggregated packets to the healthcare center according to their priorities.

Figure 5.3: Two samples of the priority calculation with *sum* and the thresholds $(t_1, t_2, \dots, t_l)$.

The proposed PPC scheme will perform user registration, data collection and WBAN-gateway packet priority classification and packet relaying steps. Fig. 5.4 depicts the data flow. The sensor-collected data are processed in the smartphone as mentioned above periodically. Then, these data are encrypted, and processed with encrypted disease model offered by the healthcare center. The calculated result will be transmitted to the WBAN-gateway. After receiving these encrypted medical packets from different users, the WBAN-gateway conducts privacy-preserving priority calculations (see Section 5.3.5) for all these medical packets, and inserts these medical packets into a priority-based relay heap. Thus, the medical packets with high priority will be relayed to the healthcare center first. In this way, the WBAN-gateway conducts the privacy-preserving priority calculation (see Section 5.3.5) based on the encrypted data in a non-interactive way, which as we know, is novel. For the reader's convenience, we summarize the important notations to be used in Table 5.1.

5.3.2 System Setup

The healthcare center is the trust entity in our model and it sets up the system, which takes a security parameter κ , and runs the bilinear generation algorithm $\mathcal{Gen}(\kappa)$ (see Section 2.2.2). The output of the generation algorithm are the public parameters $(N, g, \mathbb{G}, \mathbb{G}_T, e)$, where $\mathbb{G}, \mathbb{G}_T$ are two groups of composite order $N = pq$, p, q are two κ -bit prime numbers, g is a generator of group $\mathbb{G}$, e is a computable mapping $\mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$. Then, the healthcare center sets the BGN (see Section 2.2.3) public key $pk = (N, g, \mathbb{G}, \mathbb{G}_T, e, h)$, where $h = g^q$, and the private key $sk = p$. Moreover, the healthcare center chooses three secret random values $\alpha, \beta, \gamma \in \mathbb{Z}_N$. Next, a hybrid ECIES [116] encryption algorithm $Enc(\cdot)$ is chosen by the healthcare center. To encrypt the long user's physiological data vector and user

5.3. PROPOSED PRIVACY-PRESERVING PRIORITY CLASSIFICATION SCHEME

Figure 5.4: Data flow of the PPC scheme.

information, we choose the hybrid ECIES encryption [116] rather than a simple public encryption algorithm. Finally, the healthcare center publishes the public parameters $pk = (N, g, \mathbb{G}, \mathbb{G}_T, e, h)$ and the ECIES $Enc(\cdot)$, keeping the private parameters $(p, \alpha, \beta, \gamma)$ secret.

In order to let the WBAN-gateway conduct a privacy-preserving priority calculation over encrypted data, the healthcare center assigns the processing key $K_G = g^{\alpha p}$ to the WBAN-gateway. Next, the healthcare center defines a hash function $H(\cdot)$ and a set of hash

5.3. PROPOSED PRIVACY-PRESERVING PRIORITY CLASSIFICATION SCHEME

Table 5.1: Notations used in PPC

Notation	Description
κ	security parameter
$\mathcal{Gen}(\kappa)$	bilinear generation algorithm
$\mathbb{G}, \mathbb{G}_T$	two group of composite order $N = pq$
p, q	two κ -bit prime numbers
g	a generator of group $\mathbb{G}$
e	a computable mapping $\mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$
p	private key
$\alpha, \beta, \gamma \in \mathbb{Z}_N$	three secret random values
$Enc(\cdot)$	an hybrid ECIES [116] encryption algorithm
$K_G = g^{\alpha p}$	a processing key owned by the healthcare center
$H(\cdot)$	a hash function set by the healthcare center
$\parallel$	concatenation operation
$H_i, i \in \{1, 2, \dots, t_m\}$	a hash set created by the healthcare center
t_m	the maximum threshold for all the diseases owned by the healthcare center
$\mathbb{X} = \{x_1 \dots x_k\}$	standardized physiological data
$\vec{X}_u$	user's physiological data vector
$\vec{A} = (a_1, a_2, \dots, a_k)$	a disease model for one user
$(t_1, t_2, \dots, t_l)$	l thresholds for a user for one disease
$A_i = [a_i], i = 1, 2, \dots, k$	encrypted disease model coefficients
$T_j = [t_j], j = 1, 2, \dots, l$	encrypted thresholds

values as follows:

$$H_i = H(e(g, g)^{\alpha\beta p(\gamma+i)}), i \in \{1, 2, \dots, t_m\}$$

where t_m is the maximum threshold for all the diseases owned by the healthcare center. In other words, for each disease dis_i , according to the disease levels, it has different thresholds $t_{i1}, t_{i2}, \dots, t_{ik}$, and $t_m = \max(t_{ik})$. For example, if the healthcare center has threshold $t_{11} = 50$ for disease dis_1 , $t_{21} = 60$ for disease dis_2 and $t_{31} = 70$, $t_{32} = 65$ for disease dis_3 , then the maximum threshold $t_m = t_{31} = 70$.

For a user u , we assume that:

- The inner product of the user's physiological data vector $\vec{X}_u$ and the healthcare center's disease model $\vec{A}_u$ is positive: $\vec{X}_u \cdot \vec{A}_u > 0$;

- When the user is in a normal state for a disease level with the related threshold t_u , the inner product result is less than t_u : $\vec{X}_u \cdot \vec{A}_u < t_u$; On the other hand, if the user is in an abnormal status for threshold t_u , $\vec{X}_u \cdot \vec{A}_u \geq t_u$;

The healthcare center also sends the hash set $\{H_1, H_2, \dots, H_{t_m}\}$ and the hash function $H(\cdot)$ to the WBAN-gateway.

After the system setup, the WBAN-gateway gets the processing key $K_G = g^{\alpha p}$, the hash set $\{H_1, H_2, \dots, H_{t_m}\}$ and the hash function $H(\cdot)$.

5.3.3 User Registration

Firstly, a user registers in the healthcare center for this emergency monitoring service. As shown in the Algorithm. 4, according to the user's sensor list, the healthcare center finds the related disease that user needs to monitor. Accordingly, the healthcare center fetches the related disease model $\vec{A} = (a_1, a_2, \dots, a_k)$ and the l thresholds $(t_1, t_2, \dots, t_l)$. Note that, a user's thresholds l can be different from another user's thresholds. The healthcare center encrypts the disease model and thresholds as follows:

$$A_i = [a_i] = g^{\beta a_i} h^{r_i}, \quad i = 1, 2, \dots, k$$

$$T_j = [t_j] = g^{\beta(t_j + \gamma)} h^{r_j}, \quad j = 1, 2, \dots, l$$

where $r_i, r_j \in \mathbb{Z}_N$ are random values secretly chosen by the healthcare center. These encrypted disease model $A_i, i = 1, 2, \dots, k$ and thresholds $T_j, j = 1, 2, \dots, l$ are sent to the authorized user.

Algorithm 4 User_Register()

```

1: \ Input:  $(S_1, S_2, \dots, S_k)$ , user's sensor list
2: \ Input: userinfo, user's information like name, age, etc.
3: \ Output:  $A_i, i = 1, 2, \dots, k$ , encrypted disease model coefficients
4: \ Output:  $T_j, j = 1, 2, \dots, l$ , encrypted thresholds
5: \  $Enc(\cdot)$ , an hybrid ECIES [116] encryption algorithm
6: User does:
7:  $Enc_{sensors} = Enc(S_1|S_2|\dots|S_k|userinfo)$ 
8: send the  $Enc_{sensors}$  to the healthcare center
9:  $\longrightarrow$ 
10: Healthcare center does:
11: decrypts the  $Enc_{sensors}$ , gets the  $(S_1, S_2, \dots, S_k)$  and userinfo
12: finds out disease dis, disease model  $a_i, i = 1, 2, \dots, k$  and related thresholds  $t_j, j = 1, 2, \dots, l$  according to
    the  $(S_1, S_2, \dots, S_k)$  and userinfo
13: for  $i = 1$  to  $k$  do
14:    $A_i = [a_i] = g^{a_i} h^{r_i}$ 
15: end for
16: for  $j = 1$  to  $l$  do
17:    $T_j = [t_j] = g^{\beta(t_j + \gamma)} h^{r_j}$ 
18: end for
19: sends the  $A_i, i = 1, 2, \dots, k$  and  $T_j, j = 1, 2, \dots, l$  to the user

```

5.3.4 User Data Collection

The authorized user who is equipped with sensors collects the physiological data periodically. In each round, after sensor data standardization, the user gets the physiological data vector $\vec{X} = (x_1, x_2, \dots, x_k)$. As shown in the Algorithm. 5, with the encrypted disease model, the user performs the following computation for all the thresholds:

$$\begin{aligned}
 C'_j &= \frac{T_j}{\prod_{i=1}^k A_i^{x_i}} \cdot h^{r_j}, \quad j = 1, 2, \dots, l \\
 &= g^{\beta(\gamma + t_j - \sum_{i=1}^k a_i \cdot x_i)} \cdot h^{r_j}, \quad j = 1, 2, \dots, l
 \end{aligned}$$

where each $r_j \in \mathbb{Z}_N$ is a random value. The user also makes use of the ECIES encryption algorithm $Enc(\cdot)$ to encrypt the physiological data vector and the user's information

5.3. PROPOSED PRIVACY-PRESERVING PRIORITY CLASSIFICATION SCHEME

userinfo such as name, age, etc.

$$C = \text{Enc}(x_1||x_2||\dots||x_k||\text{userinfo})$$

Finally, the user sends the encrypted data $C||C'_1||C'_2||\dots||C'_l$ to the WBAN-gateway.

Algorithm 5 User_Data_Collect()

```
1: \ Input:  $\vec{X} = (x_1, x_2, \dots, x_k)$ , physiological data vector
2: \ Input:  $A_i, i = 1, 2, \dots, k$ , encrypted disease model coefficients
3: \ Input:  $T_j, j = 1, 2, \dots, l$ , encrypted thresholds
4: \ Output:  $C||C'_1||C'_2||\dots||C'_l$ , encrypted packet
5: user does:
6:  $prod = \prod_{i=1}^k A_i^{-x_i}$ 
7: for  $j = 1$  to  $l$  do
8:    $C'_j = T_j \cdot prod \cdot h^{r_j}$ 
9: end for
10:  $C = \text{Enc}(x_1||x_2||\dots||x_k||\text{userinfo})$ 
11: sends the  $C||C'_1||C'_2||\dots||C'_l$  to the WBAN-gateway
```

5.3.5 WBAN-Gateway Packet Classification and Relay

The WBAN-gateway receives the encrypted medical packets from different users. As shown in the Algorithm. 6, for each medical packet, the WBAN-gateway calculates the priority, and inserts it into a priority-based relay heap.

In the priority-based packet classification algorithm, the WBAN-gateway takes the $C'_1, C'_2, \dots, C'_l$ and the processing key K_G as the inputs, performs the priority calculations as follow:

$$\begin{aligned}
 D_j &= e(C'_j, K_G) \\
 &= e(g^{\beta(\gamma+t_j-\sum_{i=1}^k a_i \cdot x_i)} \cdot h^{r_j}, g^{\alpha p}) \\
 &= e(g, g)^{\alpha \beta p(\gamma+t_j-\sum_{i=1}^k a_i x_i)}
 \end{aligned}$$

where $j = 1, 2, \dots, l$, e is a computable mapping (see Section 2.2.3). The WBAN-gateway assigns the priority of the medical packet based on which hash value $H(D_j), = 1, 2, \dots, l$ is in the hashset $\{H_1, H_2, \dots, H_{l_m}\}$. Then the medical packet is inserted into the priority-based relay heap, the root packet of which is the highest priority packet. However, it is unreasonable for the low priority packets to be preempted by the high priority packets all the way. So a timer frequently increases the priority by one for all the packets in the priority-based relay heap.

5.3.6 Correctness Analysis

Firstly, we prove that for a healthy user u , the WBAN-gateway assigns the $pr_u = 1$ priority to this user's packet, if $0 < \vec{A} \cdot \vec{X} < t_1$. Recall that all the physiological data $x_i \in \mathbb{X}$ are positive, and the healthcare center's disease model coefficients $a_i, i = 1, 2, \dots, k$ are positive. It's easy to find out $\vec{A} \cdot \vec{X} = \sum_{i=1}^k a_i \cdot x_i > 0$. On the other hand, For the threshold t_1 , the

5.3. PROPOSED PRIVACY-PRESERVING PRIORITY CLASSIFICATION SCHEME

Algorithm 6 Packet_Classify_Relay()

```

1: \ Input:  $C'_1, C'_2, \dots, C'_l$ , encrypted data got from a user
2: \ Result: assigns priority  $pr_u$  to the packet, and insert it into the relay heap
3: \  $C$ :  $C = Enc(x_1 || x_2 || \dots || x_k || userinfo)$ , encrypted physiological data and user info
4: \  $K_G$ , a processing key
5: WBAN-gateway does:
6: for  $j = 1$  to  $l$  do
7:    $D_j = e(C'_j, K_G)$ 
8: end for
9:
10:  $pr_u = -1$ 
11: for  $j = 1$  to  $l$  do
12:   if  $H(D_j) \in \{H_1, H_2, \dots, H_{t_m}\}$  then
13:      $pr_u = j$ 
14:     break
15:   else if  $H(D_j) \notin \{H_1, H_2, \dots, H_{t_m}\}$  then
16:     go on
17:   end if
18: end for
19: if  $pr_u = -1$  then
20:    $pr_u = l + 1$ 
21: end if
22:  $C.priority = pr_u$ 
23:  $RelayHeap.insert(C)$ 
24:
25:  $Relaypacket = RelayHeap.removeMax()$ 
26: relay the  $Replaypacket$  to healthcare center
27: for  $i = 1$  to  $RelayHeap.size()$  do
28:    $packet = RelayHeap.packet(i)$ 
29:    $packet.priority ++$ 
30: end for

```

WBAN-gateway conducts the calculation:

$$\begin{aligned}
 D_1 &= e(C'_1, K_G) \\
 &= e(g^{\beta(\gamma+t_1-\sum_{i=1}^k a_i \cdot x_i)} \cdot h^{r_j}, g^{\alpha p}) \\
 &= e(g, g)^{\alpha \beta p(\gamma+t_1-\sum_{i=1}^k a_i x_i)}
 \end{aligned}$$

5.3. PROPOSED PRIVACY-PRESERVING PRIORITY CLASSIFICATION SCHEME

Because $t_1 > \sum_{i=1}^k a_{ui}x_{ui}$, it ensures $t_1 > t_1 - \sum_{i=1}^k a_i x_i > 0$. We define a function $PH(x) = H(e(g, g)^{\alpha\beta p(\gamma+x)})$. It is easy to find out $PH(t_1 - \sum_{i=1}^k a_i x_i)$ is in the set $\{PH(1), PH(2), \dots, PH(t_1)\}$. Therefore, $H(D_1)$ is in the hashset $\{H_1, H_2, \dots, H_{t_1}\}$, which is the subset of $\{H_1, H_2, \dots, H_{t_m}\}$. So the priority of this user's packet is $pr_u = 1$. Let $P(N, q)$ denote the probability of at least one collision when hash values are generated in this system, where N is the total number of possible hash values and q is the number of hash values this system generates. Then, according to Birthday Bound [117], we have $P(N, q) \leq q(q-1)/2N$.

Secondly, we prove that for a user u with $t_{bottom} < \sum_{i=1}^k a_i x_i < t_{up}$, the WBAN-gateway assigns up to this user's packet $pr_u = up$. Same as above, if $\sum_{i=1}^k a_i x_i < t_{up}$, $H(D_{up})$ is in the hashset $\{H_1, H_2, \dots, H_{t_{up}}\}$, which is the subset of $\{H_1, H_2, \dots, H_{t_m}\}$. For the $t_{bottom} < \sum_{i=1}^k a_i x_i$, we have the calculation:

$$\begin{aligned}
 D_{bottom} &= e(g, g)^{\alpha\beta p(\gamma+t_{bottom}-\sum_{i=1}^k a_i x_i)} \\
 \Rightarrow: t_{bottom} &< \sum_{i=1}^k a_i x_i, t_{bottom} - \sum_{i=1}^k a_i x_i < 0 \\
 PH(t_{bottom} - \sum_{i=1}^k a_i x_i) &= PH(value < 0) \\
 \Rightarrow: PH(value < 0) &\notin \{PH(1), PH(2), \dots, PH(t_m)\} \\
 PH(t_{bottom} - \sum_{i=1}^k a_i x_i) &\notin \{PH(1), PH(2), \dots, PH(t_m)\} \\
 H(D_{bottom}) &\notin \{H_1, H_2, \dots, H_{t_m}\}
 \end{aligned}$$

Thus, in the line 15-16 of the Algorithm 6, the algorithm will go on until meeting $H(D_{up})$. Then, the WBAN-gateway assigns the priority up to the user's packet, which means that

$pr_u = up$.

Thirdly, we prove that the medical packets aggregated in the WBAN-gateway will be relayed to the healthcare center according to their priorities. As mentioned above, each medical packet C containing the encrypted user physiological data $x_1, x_2, \dots, x_k$ and user information $userinfo$ is inserted into the relay heap. This heap is a max heap (mentioned in the section 2.2.4) based on the packets' priorities. The WBAN-gateway removes and relays the maximum element (root element), which is the packet with the highest priority. Then the heap is sifted to be a full tree again. Moreover, a timer frequently updates the priorities of all the packets in the heap to prevent too long starving time for low priority packets.

5.4 Security Analysis

In this section, we analyze the security properties of the proposed priority classification scheme. Specifically, following the security requirements discussed in section 5.2.2, our analysis will focus on how the proposed scheme can achieve the users' data privacy and healthcare center's disease model confidentiality.

In our system model, the users' data privacy involves protecting the sensitive physiological data in the relay packets and the procedure of priority classification. On the other hand, the healthcare center's disease model confidentiality includes protecting the disease model coefficients and the thresholds for different authenticated users. The security of the remote monitoring e-healthcare system depends on the underlying standard encryption, and our PPC scheme. The standard encryption is responsible for preventing the WBAN-gateway from learning the users' personal information. And our PPC scheme is responsible for protecting the users' data and the healthcare center's disease model while achieving the priority

classification for all the packets from different users. The ECIES encryption that consists of the symmetric encryption and the public key encryption is secure under the cipher-only, known-sample and known-plaintext attacks [116]. Thus, we focus on the analysis of the user's data privacy and the healthcare center's disease model confidentiality with our PPC scheme.

Security of the user's data encryption and the disease model encryption. First, in our PPC scheme, the coefficients of the disease model $a_i, i = 1, 2, \dots, k$ are encrypted as

$A_i = [a_i] = g^{\beta a_i} h^{r_i}, i = 1, 2, \dots, k$, and the thresholds $t_j, j = 1, 2, \dots, l$ for the user are encrypted as $T_j = [t_j] = g^{\beta(t_j + \gamma)} h^{r_j}, j = 1, 2, \dots, l$. Because the BGN encryption is secure [93], without knowing the private parameters $\{\alpha, \beta, \gamma\}$ and $sk = p$, it is hard for the WBAN-gateway to recover $a_i, i = 1, 2, \dots, k$ and $t_j, j = 1, 2, \dots, l$. Moreover, the user's physiological data $x_i, i = 1, 2, \dots, k$ are encrypted in the $C'_j = g^{\beta(\gamma + t_j - \sum_{i=1}^k a_i \cdot x_i)} \cdot h^{r_j}, j = 1, 2, \dots, l$, from which, it is hard for the WBAN-gateway to recover the $x_i, i = 1, 2, \dots, k$ directly. As mentioned above, the user's physiological data $x_i, i = 1, 2, \dots, k$ that are encrypted inside $C = Enc(x_1 || x_2 || \dots || x_k || userinfo)$ are also secure due to the security of the ECIES encryption.

Security of the user's data and healthcare center's disease model in the priority classification scheme. As discussed in the paragraph, it is hard to reveal the user's data and the healthcare center's disease model from the encrypted data. Alternatively, the attacker may attempt to recover the priority of the user's medical packet or the inner product $\vec{A} \cdot \vec{X}$ first, then reveal the vector containing the disease model and the user's physiological data. Now we show the countermeasures against this kind of attacks.

The attacker may attempt to recover the inner product of the priority calculation over $D_j, j = 1, 2, \dots, l$. However, each D_j is calculated by the WBAN-gateway with C'_j and K_G

as:

$$D_j = e(g, g)^{\alpha\beta p(\gamma + t_j - \sum_{i=1}^k a_i x_i)}$$

Because both $e(g, g)^{\alpha\beta p}$ and γ are unknown, the attacker is unable to solve $t_j - \sum_{i=1}^k a_i x_i$, according to the discrete logarithm problem. Therefore, the attacker could not recover the $a_i, i = 1, 2, \dots, k, x_i, i = 1, 2, \dots, k, t_j, j = 1, 2, \dots, l$.

Knowing the priority-based packet relay heap, the attacker may also attempt to recover the user's data and the healthcare center's disease model from the the heap. The WBAN-gateway assigns the priority of the packets based on the hash function $H(D_j)$ and the hash-set $\{H_1, H_2, \dots, H_m\}$. The hash function is non-invertible. The membership operation on whether $H(D_j)$ is a member of the hashset $\{H_1, H_2, \dots, H_m\}$ only reveals the priority of the packet, which is not a privacy concern for our system model. The timer updates the priority of the packets in the priority-based relay heap. All the operation over the heap relies on the packet's calculated priority, which reveals no user's physiological data and the healthcare center's disease model.

5.5 Performance Evaluation

In this section, we evaluate the performance of the proposed PPC scheme in terms of computational cost and communication overhead.

5.5.1 Implementation and Experimental Settings

We have implemented the PPC scheme in java. We test the PPC's performance in a testbed of a mac osx laptop and one android phone. We deploy the the programs of the healthcare center and WBAN-gateway in the mac osx laptop. The android phone plays the role of the mobile user. For comparison, we also implement a paillier-based privacy-preserving priority classification scheme. Thus, an ECS server plays the role of the decryption helper in the paillier-based scheme similar to [2]. The hardware and software of these machines are shown in Table 5.2.

Table 5.2: Experimental setting

Role	Machine	Hardware & Software
Health center	Mac laptop	CPU:2.9 GHz Intel Core i5, memory: 8GB
WBAN-gateway	Mac laptop	CPU:2.9 GHz Intel Core i5, memory: 8GB
User	MEIZU phone	CPU: Exynos 7872 3 GB ram; Android 6.0.1
Decryption helper(paillier)	Alibaba ECS	Instance: ecs.xn4.small, CentOS 7.2 64-bit and java

5.5.2 Pailliar-based privacy-preserving priority classification scheme

To clearly evaluate the performance of the proposed PPC scheme, we construct a pailliar-based privacy-preserving priority classification scheme, which is similar to the rencent work [2] for comparison. In the work [2], the authors conduct the clinical decision using the paillier encryption technique and the linear classification. As shown in Fig. 5.5, in the pailliar-based priority classification scheme, the WBAN-gateway fulfills the priority classification task for the aggregated medical packets with the help of a decryption helper, who

owns the private key $sk_{paillier}$ and the decryption algorithm $Dec(\cdot, sk_{paillier})$ of the paillier-based encryption system. The $Enc_p(\cdot)$ and $Dec_p(\cdot)$ are the encryption algorithm and the decryption algorithm of the pailliar encryption system. Specifically, in step 2, the random number r_j is a small number, which would not make the bit length of $[(t_j - \sum_{i=1}^k x_i \cdot a_i) \cdot r_j]$ change too much. Moreover, in step 6, the WBAN-gateway judges the priority of the packet according to the bit length of $[(t_j - \sum_{i=1}^k x_i \cdot a_i) \cdot r_j]$, which means that: 1) if the bit length of $[(t_j - \sum_{i=1}^k x_i \cdot a_i) \cdot r_j]$ is close to the bit length of n , t_j is less than $\sum_{i=1}^k x_i \cdot a_i$; 2) otherwise, t_j is larger than $\sum_{i=1}^k x_i \cdot a_i$.

Figure 5.5: Paillier-based priority classification scheme similar to the work [2]

5.5.3 Healthcare center's computational cost

As discussed in Section 5.3.2, in the proposed PPC scheme, the healthcare center is the trust authority who initializes the system including: setting up the BGN encryption, setting up ECIES encryption and setup the hashset $\{H_1, H_2, \dots, H_{t_m}\}$. The system initialization is conducted only once, so we focus on the healthcare center's computational cost in the user registration algorithm (Section 5.3.3).

In the user registration process, the healthcare center fetches a disease model and associated thresholds, encrypted these data. Thus, we measure the time for the healthcare center to encrypt the disease model with 50, 100, 150, 200, 250 coefficients, and 10, 20, 30, 40, 50 thresholds per disease. As shown in Fig. 5.6a and Fig. 5.6b, the encryption latency increases slowly with the number of the disease model coefficients and the thresholds. For the max number of 250 disease model coefficients and 50 thresholds, the healthcare center spends 161 milliseconds and 34 milliseconds to encrypt these data. The user registration is conducted only once for a user, so the computational cost is acceptable for both the user and the healthcare center.

We also shows the latency for disease model encryption and threshold encryption for paillier-based priority classification scheme (a competing benchmark) in Fig. 5.6. The figure shows the computational costs of healthcare center for PPC scheme and paillier-based scheme are almost the same.

5.5.4 Computational cost at user side

Assume that a user gets an encrypted disease model with k coefficients and l thresholds from the healthcare center, the user spends most of the time computing $C'_j, j = 1, 2, \dots, l$. Each $C'_j = \frac{T_j}{\prod_{i=1}^k A_i^{x_i}} \cdot h^{r_j}$ involves $k + 1$ exponentiations, $k - 1$ additions, one division and one multiplication. The algorithm can be optimized to speed up by computing the time-consuming $\prod_{i=1}^k A_i^{x_i}$ before the calculations of C'_j . According to our tests, the latency for the calculations of $C'_j, j = 1, 2, \dots, 50$ increases only with the number of $C'_j, j = 1$, so we measure the latency of user data collection for $C'_j, j = 1, 50$, which means the data collection latency for one threshold and 50 thresholds, and the results are shown in Fig. 5.7. As expected, the latency increases slowly with increasing disease dimension (number

5.5. PERFORMANCE EVALUATION

(a) Disease model encryption

(b) Threshold encryption

Figure 5.6: Computational cost of healthcare center for PPC and Paillier-based scheme [2]

of disease). The time complexity in the user side is $O(kl)$. Moreover, we measure the latency for the paillier-based priority classification scheme. As shown in Fig. 5.7, the user data collection latency of the paillier-based scheme is almost the same as the PPC scheme.

Figure 5.7: Latency for user data collection for PPC and Paillier-based scheme [2]

5.5.5 Computational cost at WBAN-gateway side

In the packet priority classification and relay algorithm (Section 5.3.5), the WBAN-gateway conducts l mapping operations for the $D_j = e(C'_j, K_G)$, $j = 1, 2, \dots, l$. After calculating the priority of the packet, the WBAN-gateway inserts the packet into the relay heap, relays the packet with the highest priority from the relay heap, and updates the priorities of the packets in the relay heap. Therefore, the latency for the WBAN-gateway consists of two parts: 1) time for packet priority classification; 2) time for inserting the packet into the relay heap and updating the relay heap. The latency for the second part is $O(d)$, in which d is the height of the relay heap. We have evaluated that this algorithm is very efficient and does

5.5. PERFORMANCE EVALUATION

not cost too much time. So we focus on the first latency. As shown in Fig. 5.8, the priority classification latency increases slowly with the number of the thresholds in the packet. Even for 50 thresholds, which is large in real scenarios, the priority classification latency is 409 milliseconds. The results demonstrate that the priority classification algorithm is very efficient in computational cost.

Figure 5.8: Latency for priority classification for PPC and Paillier-based scheme [2]

In Fig. 5.8, we also demonstrate the packet priority classification latency of the paillier-based scheme, which involves: 1) The decryption helper decrypts the C'_j , $j = 1, 2, \dots, l$; 2) The WBAN-gateway judges the priority according to $[t_j - \sum_{i=1}^k x_i \cdot a_i]$; 3) The WBAN-gateway inserts the packet into the heap and relays the packet with highest priority to the healthcare center. As we have measured, the algorithms for part 2 and part 3 do not cost too much. Thus, we focus on the latency for the part 1, which includes the communication latency for data transmission between the WBAN-gateway and the decryption helper. We measure the communication latency for the data transmission, which is around 636

milliseconds. We compare the total latency for the proposed PPC scheme with the paillier-based scheme in Fig. 5.8, which shows the priority classification latency for the paillier-based scheme is much larger than the PPC scheme because of the large communication latency between the WBAN-gateway and the decryption helper.

5.5.6 Communication Cost

Our PPC scheme is a non-interactive scheme, which ensures the scheme is very efficient in communication cost. We focus on 1) the communication cost between the healthcare center and user in user registration algorithm; 2) the communication cost between the authenticated user and the WBAN-gateway in user data collection algorithm.

In the user registration, the healthcare center sends the encrypted disease model $A_i, i = 1, 2, \dots, k$ and the encrypted thresholds $T_j, j = 1, 2, \dots, l$ to the authenticated user. As the BGN encryption is 1024 bits length in our setting, the communication cost for the healthcare center to the user is $(k + l) * 1024$ bits. This communication only occurs once for one user, so it is acceptable for a user.

In the user data collection algorithm, the user sends the encrypted $C'_j, j = 1, 2, \dots, l$ and the C to the WBAN-gateway. In our setting, the relationship of the length of plain text and cipher text in the ECIES encryption is $CipherTextLen = (PlainTextLen/BLOCKSIZE + 1) * BLOCKSIZE$. We set the block size to 16. Therefore, the communication cost between authenticated user and the WBAN-gateway in user data collection algorithm is about $1024 * l + (userInfoLen + 16)$ bits, where the $userInfoLen$ is the length of the user's original physiological data. Specifically, the communication cost for the packet relay from the WBAN-gateway to the healthcare center is $userInfoLen + 16$ bits, because only the encrypted C needs to be relayed to the healthcare center. What's more, the WBAN-gateway

achieves the privacy-preserving priority computation tasks by itself. Our PPC scheme is efficient compared with the paillier based scheme for the the additional communication costs between the WBAN-gateway and the decryption helper in pailler based scheme.

5.6 Conclusions

In this chapter, we have proposed an efficient privacy-preserving priority classification (PPC) scheme on patient healthcare data in remote e-healthcare system. The proposed PPC scheme achieves the priority classification and packets relay tasks, while preserving the privacy of the users and the confidentiality of the healthcare center's disease models. Because it is a non-interactive procedure, the communication cost is low. The evaluation demonstrates that our PPC scheme is efficient in computational cost and communication overhead. According to our tests, the latency of disease model encryption, threshold encryption and user data collection are almost the same between our PPC scheme and the paillier-based scheme. However, in the priority classification algorithm, our PPC scheme is between 6 times and 1.8 times faster than the paillier-based scheme for number of the thresholds ranging from 10 to 50.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this thesis, several privacy-preserving disease detection e-Healthcare systems are proposed to achieve the on-the-go e-Healthcare service, while preserving the security and privacy issues of the systems.

- Firstly, we have proposed an efficient and privacy-preserving pre-clinical guidance scheme (PGuide). The scheme has two key phases. Firstly, it employs an efficient privacy-preserving comparison protocol (PPCP), which enables a user to obtain disease risk predication services from a service provider without compromising the privacy of the user and the server provider. Secondly, it employs a single-attribute encryption technique to conduct an efficient privacy-preserving hospital recommendation service. Our security analysis demonstrated that the PPCP and hospital recommendation service achieve the privacy-preserving requirements. Evaluations demonstrate that at the user side, our proposed PGuide scheme's computational latency is

at most 10 ms, which is significantly smaller than the widely used paillier-based scheme's computational latency, which climbs from 120 ms to 1200 ms for the length of vector m range from 10 to 100. Moreover, at the server side, our proposed PGuide scheme is around 20 times faster than the paillier-based scheme. According to our tests, although the recommendation of hospitals in our PGuide is two times slower than AES encryption, our PGuide is more scalable than the AES encryption. Specifically, at the user side, the message encryption latency (around 2ms) of our PGuide is much smaller than the AES encryption's encryption latency, which climbs from 5 ms to 85 ms for the number of hospitals range from 50 to 250. What's more, at the user side, the key size (around 120 bytes) of our PGuide is much smaller than AES encryption's key size, which is about 1250 bytes for 80 hospitals. Moreover, the communication cost (408 bytes) of our PGuide is much smaller than AES encryption's communication cost, which climbs from 500 bytes to around 2400 bytes for the number of the hospitals ranging from 20 to 100.

- Secondly, we have proposed an efficient and privacy-preserving health query over outsourced cloud, which preserves user's personal privacy and achieves the confidentiality of the health service provider's diagnosis model. To realize HeOC, we have designed a sensor anomaly detection technique to find the high risk disease, then with oblivious pseudorandom function protocol, the user queries the health diagnosis result accurately. In addition, performance evaluation demonstrates that a user with hundreds of diseases to verify only needs around 1 second to verify and check all the information to find the disease with high risk. It is very efficient. At the cloud side, even with 10^6 diseases, the computational cost to search the anomaly leaf node is around 0.06 ms. At the cloud side, it takes $32 * n * (\log D + 1)$ bits to download

all the leaf nodes and the accompany list, where D is the total diseases number and n is the number of diseases to detect. Considering D is around 10^6 , the communication cost is acceptable.

- Thirdly, we have proposed an efficient privacy-preserving priority classification (PPC) scheme on patient healthcare data in remote e-Healthcare system. The proposed PPC scheme achieves the priority classification and packets relay tasks, while preserving the privacy of the users and the confidentiality of the healthcare center's disease models. The WBAN-gateway achieves the priority classification in a non-interactive way. According to our tests, the latency of disease model encryption, threshold encryption and user data collection are almost the same between our PPC scheme and the paillier-based scheme. However, in the priority classification algorithm, our PPC scheme is between 6 times and 1.8 times faster than the paillier-based scheme for number of the thresholds ranging from 10 to 50. The communication cost between authenticated user and the WBAN-gateway in the user data collection algorithm is about $1024 * l + (userInfoLen + 16)$ bits, where the *userInfoLen* is the length of the users original physiological data. Furthermore, the WBAN-gateway achieves the privacy-preserving priority computation tasks by itself. In contrast, the paillier based scheme incurs additional communication costs between the WBAN-gateway and the decryption helper.

6.2 Future Work

E-Healthcare Systems is a promising platform in pervasive environments. The privacy-preserving disease detection e-Healthcare systems proposed in this thesis can be extend in

6.2. FUTURE WORK

several directions. The following research topics will be investigated as a continuation of my Ph.D. thesis work.

- *Privacy-preserving disease model analysis.* In big data era, the explosive data mining techniques are used as popular tools to mine useful knowledge for the hospitals. However, considering the complexity of these techniques, the hospitals tend to out-source both data and calculations to computationally powerful cloud, which however poses a potential threat to user's privacy. Artificial neural networks (ANNs) could be beneficial as a tool for attention focusing and used to help building decision support models, which thus are able to alert both patients and healthcare practitioners of high risk of certain diseases or attributes. Tools like this can be a more cost-effective way of preliminary diagnosis that could or could not lead to additional and expensive diagnostic testing or variances in patient treatment plans. For big-data initiatives to succeed, the healthcare system must undergo some fundamental changes. For example, stakeholders across the industry need to protect patient privacy as more information becomes public, and ensure that safeguards are in place to protect organizations from releasing information. In a nutshell, the privacy issue has become a big challenge for traditional ANNs in big data era.
- *Efficient non-interactive disease detection e-Healthcare systems.* In most of the recent researches about privacy-preserving e-Healthcare systems, it involves interactive communications among multiply entities, because of the privacy-preserving computations. Although some techniques like homomorphic encryption achieve the non-interactive computations on encrypted data, the efficiency of the algorithms becomes a big issue. On the other hand, most of the privacy-preserving algorithms

are not efficient in computational cost and communication overhead in some complicated algorithms such as the algorithms for the disease model analysis. A non-interactive privacy-preserving disease detection e-Healthcare systems will make the e-Healthcare systems more practical and pervasive.

- *Differential-privacy-based e-Healthcare systems.* In some scenario, the differential privacy based techniques distort the data, making the analysis result unsuitable when high accuracy is required. It is a balance between utility and privacy. What's more, the differential privacy based technique is not scalable. Specifically, when the target model changes, the differential privacy model has to be adjusted accordingly. But in some cases, differential privacy is easy to use to implement a privacy-preserving information release system. Thus, it is meaningful to make use of the differential privacy technique to implement the e-Healthcare system, and avoid its shortcomings.

List of Author's Publications

Journal papers:

- [1] **G.Wang**, R.Lu, and Y L.Guan, “Enabling efficient and privacy-preserving health query over outsourced cloud”, *IEEE Access*, vol. PP, pp. 11, 11 2018.
- [2] **G.Wang**, R.Lu, and Y L.Guan, “Achieve privacy-preserving priority classification on patient health data in remote eHealthcare system”, *IEEE Access*, pp. 1-1, 1 2019.
- [3] **G.Wang**, R.Lu, C.Huang, Y L.Guan, “An Efficient and Privacy-Preserving Pre-Clinical Guide Scheme for Mobile eHealthcare”, *Journal of Information Security and Applications*, 46, Elsevier, pp.271-280, 6 2019.

Conference papers:

- [1] **G.Wang**, R.Lu, C.Huang, “PGuide: An Efficient and Privacy-Preserving Smartphone-Based Pre-Clinical Guidance Scheme”, *IEEE Globecom'15*, San Diego, CA, USA, December 6 - 10, 2015.
- [2] **G.Wang**, R.Lu, C.Huang, “PSLP: Privacy-Preserving Single-Layer Perceptron Learning for e-Healthcare”, *ICICS'15*, Singapore, December 2 - 4, 2015.

Bibliography

- [1] E. Ayday, J. L. Raisaro, P. J. McLaren, J. Fellay, and J. P. Hubaux, “Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data,” in *Usenix Conference on Safety, Security, Privacy and Interoperability of Health Information Technologies*, 2013, pp. 1–1.
- [2] 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 journal of biomedical and health informatics*, vol. 18, no. 1, pp. 56–66, 2014.
- [3] Q. A. MARILYN LINTON. Doctor’s office waiting times increasingly frustrating. [Online]. Available: <http://www.torontosun.com/2011/08/17/doctors-office-waiting-times-increasingly-frustrating>
- [4] E. HUANG. It isn’t getting any easier to get a doctor’s appointment in china. [Online]. Available: <http://www.theatlantic.com/china/archive/2013/05/it-isnt-getting-any-easier-to-get-a-doctors-appointment-in-china/276400/>
- [5] jessie. Waiting all night outside a hospital hoping to see a doctor. [Online]. Available: <http://www.chinasmack.com/2009/pictures/chinese-waiting-hospital-crowds.html>

- [6] ——. List of different types of doctors and what they do. [Online]. Available: <http://mynamein.wordpress.com/2011/03/28/list-of-different-types-of-doctors-and-what-they-do/>
- [7] G. J. Annas *et al.*, “Hipaa regulations-a new era of medical-record privacy?” *New England Journal of Medicine*, vol. 348, no. 15, pp. 1486–1490, 2003.
- [8] J. Winkler, R. Ehret, T. Büttner, U. Dillmann, W. Fogel, M. Sabolek, J. Winkelmann, and J. Kassubek, “Parkinson disease risk score: moving to a premotor diagnosis,” *Journal of neurology*, vol. 258, no. 2, pp. 311–315, 2011.
- [9] Y. Wang, L. Sun, and J. Hou, “Hierarchical medical system based on big data and mobile internet: a new strategic choice in health care,” *JMIR medical informatics*, vol. 5, no. 3, p. e22, 2017.
- [10] R. Gravina, A. Alessandro, A. Salmeri, L. Buondonno, N. Raveendranathan, V. Loseu, R. Giannantonio, E. Seto, and G. Fortino, “Enabling multiple bsn applications using the spine framework.” in *International Conference on Body Sensor Networks*, 2010, pp. 228–233.
- [11] F. Bellifemine, G. Fortino, R. Giannantonio, R. Gravina, A. Guerrieri, and M. Sgroi, “Spine: a domain-specific framework for rapid prototyping of wbsn applications,” *Software-practice & Experience*, vol. 41, no. 3, pp. 237–265, 2011.
- [12] G. Fortino, R. Giannantonio, R. Gravina, P. Kuryloski, and R. Jafari, “Enabling effective programming and flexible management of efficient body sensor network applications,” *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 1, pp. 115–133, 2012.

- [13] C. Dwork and M. Naor, “On the difficulties of disclosure prevention in statistical databases or the case for differential privacy,” *Journal of Privacy & Confidentiality*, 2008.
- [14] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, “Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing,” in *Usenix Conference on Security Symposium*, 2014, pp. 17–32.
- [15] B. Dan and M. Zhandry, “Multiparty key exchange, efficient traitor tracing, and more from indistinguishability obfuscation,” in *Cryptology Conference*, 2014, pp. 480–499.
- [16] M. V. Dijk, C. Gentry, S. Halevi, and V. Vaikuntanathan, *Fully Homomorphic Encryption over the Integers*. Springer Berlin Heidelberg, 2010.
- [17] M. Alrige and S. Chatterjee, *Toward a Taxonomy of Wearable Technologies in Healthcare*. Springer International Publishing, 2015.
- [18] C. Wang, W. Lu, M. R. Narayanan, S. J. Redmond, and N. H. Lovell, “Low-power technologies for wearable telecare and telehealth systems: A review,” *Biomedical Engineering Letters*, vol. 5, no. 1, pp. 1–9, 2015.
- [19] P. Jiang, J. Winkley, C. Zhao, R. Munnoch, G. Min, and L. T. Yang, “An intelligent information forwarder for healthcare big data systems with distributed wearable sensors,” *IEEE Systems Journal*, vol. 10, no. 3, pp. 1147–1159, 2016.
- [20] G. Wang, R. Lu, and C. Huang, “Pguide: An efficient and privacy-preserving smartphone-based pre-clinical guidance scheme,” in *IEEE Global Communications Conference*, 2015, pp. 1–6.

- [21] G. Wang, R. Lu, and Y. Guan, “Enabling efficient and privacy-preserving health query over outsourced cloud,” *IEEE Access*, vol. PP, pp. 1–1, 11 2018.
- [22] —, “Achieve privacy-preserving priority classification on patient health data in remote ehealthcare system,” *IEEE Access*, pp. 1–1, 2019.
- [23] J. Wang, Y. Zhao, S. Jiang, and J. Le, “Providing privacy preserving in cloud computing,” in *International Conference on Human System Interaction*, 2010, pp. 213–216.
- [24] H. Aamot, C. D. Kohl, D. Richter, and P. Knaup-Gregori, “Pseudonymization of patient identifiers for translational research,” *BMC Medical Informatics and Decision Making*, 13,1(2013-07-24), vol. 13, no. 1, pp. 75–75, 2013.
- [25] T. Neubauer and J. Heurix, “A methodology for the pseudonymization of medical data,” *International Journal of Medical Informatics*, vol. 80, no. 3, pp. 190–204, 2011.
- [26] T. Neubauer and A. Ekelhart, “An evaluation of technologies for the pseudonymization of medical data,” in *ACM Symposium on Applied Computing*, 2009, pp. 857–858.
- [27] J. Heurix and T. Neubauer, *Privacy-Preserving Storage and Access of Medical Data through Pseudonymization and Encryption*. Springer Berlin Heidelberg, 2011.
- [28] B. Riedl, T. Neubauer, G. Goluch, O. Boehm, G. Reinauer, and A. Krumboeck, “A secure architecture for the pseudonymization of medical data,” in *International Conference on Availability, Reliability and Security*, 2007, pp. 318–324.

- [29] M. Hansen, P. Berlich, J. Camenisch, S. Clau, A. Pfitzmann, and M. Waidner, “Privacy-enhancing identity management,” *Information Security Technical Report*, vol. 9, no. 1, pp. 35–44, 2004.
- [30] M. Shin, S. Yoo, K. H. Lee, and D. Lee, “Electronic medical records privacy preservation through k-anonymity clustering method,” in *Joint International Conference on Soft Computing and Intelligent Systems*, 2013, pp. 1119–1124.
- [31] P. Belsis and G. Pantziou, “A k -anonymity privacy-preserving approach in wireless medical monitoring environments,” *Personal & Ubiquitous Computing*, vol. 18, no. 1, pp. 61–74, 2014.
- [32] J. J. Panackal, A. S. Pillai, and V. N. Krishnachandran, “Disclosure risk of individuals: A k-anonymity study on health care data related to indian population,” in *International Conference on Data Science & Engineering*, 2014, pp. 200–205.
- [33] Y. Xie, Q. He, D. Zhang, and X. Hu, “Medical ethics privacy protection based on combining distributed randomization with k-anonymity,” in *International Congress on Image and Signal Processing*, 2016, pp. 1577–1582.
- [34] D. Wei, K. N. Ramamurthy, and K. R. Varshney, “Health insurance market risk assessment: Covariate shift and k-anonymity,” in *SIAM Data Mining*, 2015.
- [35] B. B. Mehta and U. P. Rao, “Privacy preserving big data publishing: a scalable k-anonymization approach using mapreduce,” *Iet Software*, vol. 11, no. 5, pp. 271–276, 2017.

- [36] D. Choi, D. Kim, and S. Park, “A framework for context sensitive risk-based access control in medical information systems,” *Computational and Mathematical Methods in Medicine*, 2015, (2015-5-4), vol. 2015, p. 265132, 2015.
- [37] S. Z. R. Rizvi, P. W. L. Fong, J. Crampton, and J. Sellwood, “Relationship-based access control for an open-source medical records system,” in *The ACM Symposium*, 2015, pp. 113–124.
- [38] E. Kamateri, E. Kalampokis, E. Tambouris, and K. Tarabanis, “The linked medical data access control framework,” *Journal of Biomedical Informatics*, vol. 50, no. 8, pp. 213–225, 2014.
- [39] L. Seitz, J. M. Pierson, and L. Brunie, “Semantic access control for medical applications in grid environments,” *Lecture Notes in Computer Science*, vol. 2790, pp. 374–383, 2017.
- [40] P. Vimalachandran, H. Wang, and Y. Zhang, “Securing electronic medical record and electronic health record systems through an improved access control,” vol. 9085, pp. 17–30, 2015.
- [41] H. A. Maw, H. Xiao, B. Christianson, and J. A. Malcolm, “An evaluation of break-the-glass access control model for medical data in wireless sensor networks,” in *IEEE International Conference on E-Health Networking, Applications and Services*, 2015, pp. 130–135.
- [42] J. Sun, Y. Fang, and X. Zhu, *Privacy and emergency response in e-healthcare leveraging wireless body sensor networks*. IEEE Press, 2010.

- [43] L. Chen and D. B. Hoang, “Novel data protection model in healthcare cloud,” in *IEEE International Conference on High PERFORMANCE Computing and Communications*, 2011, pp. 550–555.
- [44] H. A. J. Narayanan and M. H. Gunes, “Ensuring access control in cloud provisioned healthcare systems,” in *Consumer Communications and NETWORKING Conference*, 2011, pp. 247–251.
- [45] M. Kim and K. Lauter, “Private genome analysis through homomorphic encryption,” *Bmc Medical Informatics & Decision Making*, vol. 15, no. Suppl 5, pp. S3–S3, 2015.
- [46] J. Shi, J. Lai, Y. Li, R. H. Deng, and J. Weng, “Authorized keyword search on encrypted data,” in *European Symposium on Research in Computer Security*, 2014, pp. 419–435.
- [47] J. W. Bos, K. Lauter, and M. Naehrig, “Private predictive analysis on encrypted medical data,” *Journal of Biomedical Informatics*, vol. 50, no. 8, pp. 234–243, 2014.
- [48] Y. Liu, X. Qu, and G. Xin, “A roi-based reversible data hiding scheme in encrypted medical images ?” *Journal of Visual Communication & Image Representation*, vol. 39, pp. 51–57, 2016.
- [49] S. Wang, J. Li, C. Zhang, J. Liang, and Z. Wang, “A robust algorithm of encrypted medical volume data retrieval based on 3d dwt and 3d dft,” in *IEEE International Conference on Software Engineering Research, Management and Applications*, 2017, pp. 143–149.

- [50] M. Huang, W. Xie, and P. Zhang, “Efficient fuzzy keyword search over encrypted medical and health data in hybrid cloud,” *Journal of Medical Imaging & Health Informatics*, vol. 7, no. 4, pp. 867–874, 2017.
- [51] N. Kittawi and A. Al-Haj, “Reversible data hiding in encrypted images,” in *International Conference on Information Technology*, 2017, pp. 808–813.
- [52] Y. Elmehdwi, B. K. Samanthula, and W. Jiang, “Secure k-nearest neighbor query over encrypted data in outsourced environments,” pp. 664–675, 2014.
- [53] J. Groth, “A verifiable secret shuffle of homomorphic encryptions,” in *International Workshop on Theory and Practice in Public Key Cryptography: Public Key Cryptography*, 2003, pp. 145–160.
- [54] B. Dan and D. M. Freeman, “Homomorphic signatures for polynomial functions,” in *International Conference on the Theory and Applications of Cryptographic Techniques*, 2011, pp. 149–168.
- [55] Gentry and Craig, “Fully homomorphic encryption using ideal lattices,” *Stoc*, vol. 9, no. 4, pp. 169–178, 2009.
- [56] T. El Gamal, “A public key cryptosystem and a signature scheme based on discrete logarithms,” *IEEE Trans.inf.theory*, vol. 31, no. 4, pp. 469–472, 1984.
- [57] P. Paillier, “Public-key cryptosystems based on composite degree residuosity classes,” in *Advances in Cryptology - EUROCRYPT '99, International Conference on the Theory and Application of Cryptographic Techniques, Prague, Czech Republic, May 2-6, 1999, Proceeding*, 1999, pp. 223–238.

- [58] O. Kocabas, T. Soyata, J. P. Couderc, and M. Aktas, “Assessment of cloud-based health monitoring using homomorphic encryption,” in *IEEE International Conference on Computer Design*, 2013, pp. 443–446.
- [59] O. Kocaba and T. Soyata, *Medical Data Analytics in the Cloud Using Homomorphic Encryption*, 2016.
- [60] X. Wang and Z. Zhang, “Data division scheme based on homomorphic encryption in wsns for health care,” *Journal of Medical Systems*, vol. 39, no. 12, p. 188, 2015.
- [61] X. Yi, Y. Miao, E. Bertino, and J. Willemson, “Multiparty privacy protection for electronic health records,” in *GLOBECOM 2013 - 2013 IEEE Global Communications Conference*, 2013, pp. 2730–2735.
- [62] O. Kocabas and T. Soyata, “Towards privacy-preserving medical cloud computing using homomorphic encryption,” *Enabling Real-Time Mobile Cloud Computing through Emerging Technologies*, 2015.
- [63] X. Yi, R. Paulet, E. Bertino, and V. Varadharajan, “Practical k nearest neighbor queries with location privacy,” in *IEEE International Conference on Data Engineering*, 2014, pp. 640–651.
- [64] R. Paulet, M. G. Koasar, X. Yi, and E. Bertino, “Privacy-preserving and content-protecting location based queries,” *IEEE Transactions on Knowledge & Data Engineering*, vol. 26, no. 5, pp. 1200–1210, 2014.
- [65] X. Yi, R. Paulet, E. Bertino, and V. Varadharajan, “Practical approximate k nearest neighbor queries with location and query privacy,” *IEEE Transactions on Knowledge & Data Engineering*, vol. 28, no. 6, pp. 1546–1559, 2016.

- [66] R. Lu, K. Heung, A. Lashkari, and A. Ghorbani, "A lightweight privacy-preserving data aggregation scheme for fog computing-enhanced iot," *IEEE Access*, vol. PP, no. 99, pp. 1–1, 2017.
- [67] W. Guo, J. Shao, R. Lu, Y. Liu, and A. A. Ghorbani, "A privacy-preserving online medical prediagnosis scheme for cloud environment," *IEEE Access*, vol. PP, no. 99, pp. 1–1.
- [68] A. Page, O. Kocabas, S. Ames, and M. Venkitasubramaniam, "Cloud-based secure health monitoring: Optimizing fully-homomorphic encryption for streaming algorithms," in *GLOBECOM Workshops*, 2014, pp. 48–52.
- [69] N. McDonald, D. Atkinson, Y. Khmelevsky, and S. Mcmillan, "Sport wearable biometric data encrypted emulation and storage in cloud," in *Electrical and Computer Engineering*, 2016, pp. 1–4.
- [70] M. Li, W. Lou, and K. Ren, "Data security and privacy in wireless body area networks," *IEEE Wireless Communications*, vol. 17, no. 1, pp. 51–58, 2016.
- [71] Z. Chen, H. Hu, and J. Yu, "Privacy-preserving large-scale location monitoring using bluetooth low energy," in *International Conference on Mobile Ad-Hoc and Sensor Networks*, 2016, pp. 69–78.
- [72] C. Y. Chou, E. J. Chang, H. T. Li, and A. Y. Wu, "Low-complexity privacy-preserving compressive analysis using subspace-based dictionary for ecg telemonitoring system," *IEEE Transactions on Biomedical Circuits & Systems*, vol. PP, no. 99, pp. 1–11, 2018.

- [73] E. Ayday, J. L. Raisaro, P. J. McLaren, J. Fellay, and J. Hubaux, “Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data,” in *2013 USENIX Workshop on Health Information Technologies, HealthTech '13, Washington, D.C., August 12, 2013*, 2013.
- [74] S. Katzenbeisser and M. Petković, “Privacy-preserving recommendation systems for consumer healthcare services,” in *Third International Conference on Availability, Reliability and Security, 2008. ARES 08.* IEEE, 2008, pp. 889–895.
- [75] D. S. A. Elminaam, H. M. Abdual-Kader, and M. M. Hadhoud, “Evaluating the performance of symmetric encryption algorithms,” *International Journal of Network Security*, vol. 10, no. 3, pp. 216–222, 2010.
- [76] L. Wasserman and S. Zhou, “A statistical framework for differential privacy,” *Publications of the American Statistical Association*, vol. 105, no. 489, pp. 375–389, 2010.
- [77] F. Yu, M. Rybar, C. Uhler, and S. E. Fienberg, “Differentially-private logistic regression for detecting multiple-snp association in gwas databases,” in *International Conference on Privacy in Statistical Databases*, 2014, pp. 170–184.
- [78] F. Tram, Z. Huang, J. P. Hubaux, and E. Ayday, “Differential privacy with bounded priors:reconciling utility and privacy in genome-wide association studies,” in *ACM Sigsac Conference on Computer and Communications Security*, 2015, pp. 1286–1297.
- [79] S. Sean and B. Bonnie, “Realizing privacy preserving genome-wide association studies:,” *Bioinformatics*, vol. 32, no. 9, pp. 1293–1300, 2016.

- [80] S. Simmons, C. Sahinalp, and B. Berger, “Enabling privacy-preserving gwass in heterogeneous human populations.” *Cell Syst*, vol. 3, no. 1, pp. 54–61, 2016.
- [81] S. Wang, N. Mohammed, and R. Chen, “Differentially private genome data dissemination through top-down specialization,” *Bmc Med Inform Decis Mak*, vol. 14, no. S1, p. S2, 2014.
- [82] C. Uhlerop, A. Slavkovi?, and S. E. Fienberg, “Privacy-preserving data sharing for genome-wide association studies.” *Journal of Privacy & Confidentiality*, vol. 5, no. 1, p. 137, 2013.
- [83] D. Froelicher, P. Egger, J. S. Sousa, J. L. Raisaro, Z. Huang, C. Mouchet, B. Ford, and J. P. Hubaux, “Unlynx: A decentralized system for privacy-conscious data sharing,” *Proceedings on Privacy Enhancing Technologies*, vol. 2017, no. 4, 2017.
- [84] M. M. A. Aziz, R. Ghasemi, M. Waliullah, and N. Mohammed, “Aftermath of bus-tamante attack on genomic beacon service,” *Bmc Medical Genomics*, vol. 10, no. Suppl 2, p. 43, 2017.
- [85] S. S. Shringarpure and C. D. Bustamante, “Privacy risks from genomic data-sharing beacons,” *American Journal of Human Genetics*, vol. 97, no. 5, pp. 631–46, 2015.
- [86] P. Anooj, “Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules,” *Journal of King Saud University-Computer and Information Sciences*, vol. 24, no. 1, pp. 27–40, 2012.
- [87] X. Liu, R. Lu, J. Ma, L. Chen, and B. Qin, “Privacy-preserving patient-centric clinical decision support system on naive bayesian classification,” *Journal of Biomedical and Health Informatics*, to appear.

- [88] K. Y. Yigzaw and J. G. Bellika, “A communicable disease prediction benchmarking platform,” in *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2014. IEEE, 2014, pp. 564–568.
- [89] J. Sheng, F. Li, and S. Wong, “Optimal drug prediction from personal genomics profiles,” *Journal of Biomedical and Health Informatics*, vol. 19, no. 4, pp. 1264–1270, 2015.
- [90] W. Bouwmeester, J. W. Twisk, T. H. Kappen, W. A. Klei, K. G. Moons, and Y. Vergouwe, “Prediction models for clustered data: comparison of a random intercept and standard regression model,” *BMC medical research methodology*, vol. 13, no. 1, p. 19, 2013.
- [91] F. Zhang, R. Safavi-Naini, and W. Susilo, “An efficient signature scheme from bilinear pairings and its applications,” in *Public Key Cryptography–PKC 2004*. Springer Berlin Heidelberg, 2004, pp. 277–290.
- [92] D. Boneh, “Bilinear groups of composite order,” in *International Conference on Pairing-Based Cryptography*. Springer, 2007, pp. 1–1.
- [93] B. Dan, E. J. Goh, and K. Nissim, “Evaluating 2-dnf formulas on ciphertexts,” in *International Conference on Theory of Cryptography*, 2005, pp. 325–341.
- [94] M. D. Atkinson, J.-R. Sack, N. Santoro, and T. Strothotte, “Min-max heaps and generalized priority queues,” *Communications of the ACM*, vol. 29, no. 10, pp. 996–1000, 1986.

- [95] M. A. Soliman, I. F. Ilyas, and K. C.-C. Chang, “Probabilistic top-k and ranking-aggregate queries,” *ACM Transactions on Database Systems (TODS)*, vol. 33, no. 3, p. 13, 2008.
- [96] N. Döttling, R. Dowsley, J. Muller-Quade, and A. C. A. Nascimento, “A cca2 secure variant of the mceliece cryptosystem,” *IEEE Transactions on Information Theory*, vol. 58, no. 10, pp. 6672–6680, 2012.
- [97] K. J. Cios and G. William Moore, “Uniqueness of medical data mining,” *Artificial intelligence in medicine*, vol. 26, no. 1, pp. 1–24, 2002.
- [98] J. Daemen and V. Rijmen, *The design of Rijndael: AES-the advanced encryption standard*. Springer Science & Business Media, 2013.
- [99] D. Delen, G. Walker, and A. Kadam, “Predicting breast cancer survivability: a comparison of three data mining methods,” *Artificial intelligence in medicine*, vol. 34, no. 2, pp. 113–127, 2005.
- [100] M. Brameier and W. Banzhaf, “A comparison of linear genetic programming and neural networks in medical data mining,” *IEEE Transactions on Evolutionary Computation*, vol. 5, no. 1, pp. 17–26, 2001.
- [101] Nate. (2013) Top 10 most common diseases found in hospitals. [Online]. Available: <http://www.nursingschoolhub.com/top-10-most-common-diseases-found-in-hospitals/>
- [102] A. De Caro and V. Iovino, “jpbcc: Java pairing based cryptography,” in *Proceedings of the 16th IEEE Symposium on Computers and Communications, ISCC 2011*.

- Kerkyra, Corfu, Greece, June 28 - July 1: IEEE, 2011, pp. 850–855. [Online]. Available: <http://gas.dia.unisa.it/projects/jpbc/>
- [103] D. Nevedrov, “Using jmeter to performance test web services,” *Published on dev2dev*, 2006.
- [104] M. Naehrig, K. Lauter, and V. Vaikuntanathan, “Can homomorphic encryption be practical?” in *Proceedings of the 3rd ACM workshop on Cloud computing security workshop*. ACM, 2011, pp. 113–124.
- [105] R. Cramer, G. Hanaoka, D. Hofheinz, H. Imai, E. Kiltz, R. Pass, A. Shelat, and V. Vaikuntanathan, “Bounded cca2-secure encryption,” in *International Conference on the Theory and Application of Cryptology and Information Security*. Springer, 2007, pp. 502–518.
- [106] D. Cash, J. Jaeger, S. Jarecki, C. S. Jutla, H. Krawczyk, M. Rosu, and M. Steiner, “Dynamic searchable encryption in very-large databases: Data structures and implementation,” in *21st Annual Network and Distributed System Security Symposium, NDSS 2014, San Diego, California, USA, February 23-26, 2014*, 2014.
- [107] O. Goldreich, S. Goldwasser, and S. Micali, “How to construct random functions,” *Journal of the Acm*, vol. 33, no. 4, pp. 792–807, 1986.
- [108] WHO. List of disease categories in icd-10. [Online]. Available: <http://apps.who.int/classifications/icd10/browse/2010/en>
- [109] C. A. Otto, E. Jovanov, and A. Milenkovic, “A wban-based system for health monitoring at home,” in *Ieee/embs International Summer School on Medical Devices and BIOSENSORS*, 2006, pp. 20–23.

- [110] O. Omeni, A. Wong, A. J. Burdett, and C. Toumazou, “Energy efficient medium access protocol for wireless medical body area sensor networks,” *IEEE Transactions on Biomedical Circuits & Systems*, vol. 2, no. 4, p. 251, 2008.
- [111] A. Argyriou, A. C. Breva, and M. Aoun, “Optimizing data forwarding from body area networks in the presence of body shadowing with dual wireless technology nodes,” *Mobile Computing IEEE Transactions on*, vol. 14, no. 3, pp. 632–645, 2015.
- [112] S. Rezvani and S. A. Ghorashi, “Context aware and channel-based resource allocation for wireless body area networks,” *Iet Wireless Sensor Systems*, vol. 3, no. 1, pp. 16–25, 2013.
- [113] D. Harrison, S. Boyce, P. Loughnan, P. Dargaville, H. Storm, and L. Johnston, “Skin conductance as a measure of pain and stress in hospitalised infants,” *Early Human Development*, vol. 82, no. 9, pp. 603–608, 2006.
- [114] P. Anooj, “Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules,” *Journal of King Saud University-Computer and Information Sciences*, vol. 24, no. 1, pp. 27–40, 2012.
- [115] J. B. G. Jr, “Regression analysis and forecasting models,” *Introduction to Financial Forecasting in Investment Analysis*, pp. 277–301, 2007.
- [116] A. Ara, M. Al-Rodhaan, T. Yuan, and A. Al-Dhelaan, “A secure privacy-preserving data aggregation scheme based on bilinear elgamal cryptosystem for remote health monitoring systems,” *IEEE Access*, vol. 5, no. 99, pp. 12 601–12 617, 2017.

- [117] W. Landecker, T. Shrimpton, and R. S. Terashima, “Tweakable blockciphers with beyond birthday-bound security,” in *Annual Cryptology Conference*. Springer, 2012, pp. 14–30.