PRIVACY PRESERVATION IN MEDICAL DATASET

Computer Science & Engineering

Under the Supervision of

Dr. Pardeep Kumar/

Dr.Yugal Kumar

By

Rajat Singh

192202

Jaypee University of Information Technology

Waknaghat, Solan– 173234, Himachal Pradesh

Certificate

This is to certify that projectreport entitled “**Privacy preservation in medical dataset**.”, submitted by **Rajat Singh** in partial fulfillment for the award of degree ofMaster of Technology in Computer Science & Engineering to Jaypee University of Information Technology, Waknaghat, Solan has been made under my supervision. This report has not been submitted partially or fully to any other University or Institute for the award of this or any other degree or diploma.

Pardeep Kumar.

Date: 3rd July, 2021 Dr. Pardeep Kumar

Designation

YUGAL

Date: 3rd July, 2021

Dr.Yugal Kumar

Designation

Acknowledgemet

First and foremost I am extremely grateful to my supervisors; Prof. Dr. Pardeep Kumar and Dr.Yugal Kumar for their invaluable advice, continuous support, and patience during my MTech.Study.Their immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. I would also like to thank Dr. Vivek Kumar Sehgal and Dr. Rajinder Sandhu for their technical support on my study. It is their kind help and support that have made my study and life a wonderful time. Finally, I would like to express my gratitude to my parents. Without their tremendous understanding and encouragement in the past few years, it would be impossible for me to complete my study.

Rajat Singh

Date: 3rd July, 2021

Signature:

Rajat Singh

Contents

List of Figures

S.No. Title Page No.

List of Tables

List of Abbrivations

AES	Advanced Encryption Standard
BSN	Body Sensor Network
CDH	Computational Die-Hellman
DDH	Decisional Die-Hellman
FHE	Fully Homomorphic Encryption
HeOC	Health Query over Outsourced Cloud
HIPAA	Health Insurance Portability and Accountability Act
HIS	Hospital Information System
JPBC	Java Pairing-Based Cryptography Library
OPRF	Oblivious Pseudorandom Function Protocol
OR	Odds Ratio
PHI	Protected Health Information
PKE	Public-Key Encryption
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

Abstract

The online medical prediagnosis sysyem has emerged to easethe shortage of health professionals in rural areas. Electronic medical records (EMRs) play a critical role in health-care networks. Because these records always contain potentially sensitive information about patients, privacy preservation for the EMR system is essential.

CurrentlyStandard calculations for affiliation rule mining depend on ID of incessant clinical items. In this project, we will concentrate on how we can keep up security in a dispersed mining of the frequent clinical itemsets. That is, we concenrate how (at least two) parties discover contiguous itemsets in a disseminated data set without uncovering each gathering's piece of information to the next. Then the current answer for an upward divided information releases a lot of data, while the current answer for evenly parceled information just works for three gatherings or more. In this project, we plan calculations for in an upward direction and evenly parceled information, separately.

We give two calculations for in an upward direction parceled information: one of them uncovers just the help tally and other uncovers nothing. The two eventually of them have operational operating costs directly in quantity of exchanges.Our calculation for an evenly divided information worked for gatheringsand it is more productive than the current arrangement.

Lately, with the unstable improvement in Internet, data storage, and information management innovations, One of the more prominent worries in information mining has been protection conservation .Choice emotionally supportive networks (DSS) are intelligent based frameworks characterised to help leaders utilise information and models to distinguish issues, tackle issues and decide. Clinical choice emotionally supportive network (CDSS) is intended to help doctors and other wellbeing experts with dynamic errands, for example, deciding conclusion of patient information which improve viability. Presently a-days distant re-appropriating is utilized to diminish the weight of clinical choice help in medical services. The clinicians can utilize the wellbeing information situated in distant workers by means of the web to analyze their patients. However, these workers are outsider and in this manner conceivably not completely trusted, raise conceivable security concerns. Subsequently a security protecting convention is worked for analysis of patient information where both worker and the customer are uninformed of the interior.

CHAPTER 1

Introduction

1.1 Healthcare systems

1.1.1 Background

The lack of medical professionals, which is exacerbated by the rising number of patients and an ageing population, is one of the most critical concerns affecting hospital administration. This has led in longer wait times for patients in numerous nations. In a hospital, for example, a patient may have to wait anywhere from 4 to 24 hours to be seen. In a typical hospital, for example, there are roughly 60 distinct types of medical experts; consulting a doctor who is not certified or specialised in the particular disease would be an unpleasant and time-consuming experience. Some major hospitals have developed a manned inquiry counsellor to assist patients in making the best choices possible. A patient, for example, is unlikely or hesitant to disclose sensitive health-related topics such as HIV and mental illness in public or over the automated inquiry counters. As a result, we need a strategy to assist relieve the pressure on healthcare institutions' limited services. However, as our society becomes more digital and ubiquitous gadgets become more ubiquitous, security and privacy considerations must be considered while developing privacy-preserving healthcare facilities or e-Healthcare systems. But the point is, how can we be confident that our medical data will not be compromised if we don't provide our consent?

Is it true that third-party entities, such as insurance companies and possible employers, are not permitted to access and utilise sensitive medical data? In addition to the privacy concerns of medical customers, information leakage is a key problem for service providers, especially in nations where medical providers are subject to strict regulatory regimes, such as the United States.

1.1.2 Applications of e-Healthcare

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 consumers visit the specialists, they may utilise their health data and other sources to make a preliminary diagnostic, and then obtain recommendations for therapy. Using this primeliquid diagnostic method can assist with both the heuristic and hierarachical medical systems.

- **Elder-assistance:**When an old person is matched with sensitive sensors, he or she can obtain menstrual monitoring services from the Healthy Center. There is a chance that your smart phone, smart watch, or some other gadget may be able to collect sensory data and carry out diagnostic. The elder-assistant system can detect the health condition of the older person and alert you if accidents are possible.

1.1.3 Challenges for e-Healthcare system

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

- **Challenges to privacy and security.**The majority of the users' physiological data, sensitive information, and the condition patterns of the healthcare centre must be encoded. An attacker should not be able to decrypt critical information just by looking at the encrypted text. Furthermore, it is plausible to presume that the attacker is aware of specific users' information or suffers from an illness in specific cases.

- **Challenges to accuracy.**Personal information and the illness profile of the health care centre have to be standardised to fulfil the security standards of the remote e-Health system. Using the typical benchmarking techniques may jeopardise the accuracy of the computation. As well, several questions regarding regularisation and computation efficiency remain.

- **Computational efficiencyc hallenges.** Most privacy-preserving techniques based on encryption have a large computational cost. Most privacy-preserving techniques have too many participants, and as a result, their communication costs are considerable. Also, the non-interactive confidential systems commonly use long procedures in conjunction with secrecy systems that are not interactive..

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.**Each customer uses the cloud to conduct health inquiries, gathering physiologically and personally identifiable information, such as age, gender, and medical records. Keep this personal data secure and separate from the cloud. User physiogonomic data should not be accessible to the attacker in any way.

•**Confidentiality.** Because these models constitute the hospital's intellectual properties, they must be protected from being disseminated to the general public. As previously stated, the cloud should not be able to recover the diagnostic model using protected medical data obtained from the hospital, suggesting that the system should be safe.

•**Authentication.** This sentence means "It's a statement on how nicely something is done." Only those authorised may send in queries on health and receive responses on disease. In the event that an unauthorised user performs a health-related enquiry, it should be brought to the attention promptly..

1.2 The Purposes and Goals of Research

1.2.1 Motivations

The following are the motivations for this research:

• **User's privacy and system security in two-party online health query system**. A new age of open data is rising in the realm of health care. In order to acquire new insights, healthcare experts, service providers, and other field experts have made use of voluminous amounts of data. despite the fact that these new initiatives are only at the beginning stages, they may provide assistance to the industry by contributing to the understanding of variances in health care quality issues To learn which treatments are the most effective for particular ailments, scientists might mine past data for information, supplying crucial information that could help patients and save money. Personal information for each user is shared via a two-party online healthcare inquiry system that connects them to the health facility. Personal information would be encoded first to ensure privacy and security. The healthcare centre must utilise the encrypted personal information of the user to produce an ailment diagnosis.

• **User's Privacy and data confidentiality in cloud based system.** The user's privacy and the healthcare center's data confidentiality are both safeguarded in a cloud-based online health inquiry system. The cloud-based remote e-Healthcare system has gotten a lot of attention and is getting more popular as cellphones, wireless body area networks, and cloud computing become more common. In recent years, a number of BSN applications have been proposed, including. The healthcare centre outsources illness diagnosis responsibilities to the cloud service provider in the cloud-based online health inquiry system. Each customer gives the cloud provider his or her sensibly gathered data. The user's privacy and the healthcare center's data confidentiality should be secured because the cloud service provider is only semi-trustworthy.

• **Computational efficiency and user's privacy in information relay system**. In certain information relay systems, the relay channel is semi-trust. Because some medical systems, such as the elderly-assistance system, are in an emergency, the relay channel would reply first and foremost to medical packets of high severity. As a result, both computing efficiency and privacy concerns for users must be addressed.

1.2.2 Objectives

As previously said, e-Healthcare systems address security and privacy problems, hence research into a privacy-preserving e-Healthcare system appears promising. It is possible to develop a range of privacy-preserving disease detection methods and data analysis remote e-Healthcare systems. Among these, more difficult security and privacy concerns are also reared up according to system models. For example, in certain e-Healthcare systems, a semi-honest service provider (cloud) is involved to make security and privacy concerns more complicated.

• In certain two-party system models, which only include users and a healthcare centre, each user sends a health inquiry to the healthcare centre. We focus on building a two-party online inquiry system between consumers and a healthcare centre in this scenario, while preserving the user's personal information, privacy, and system security.

- In certain cloud-based multi-party system models, which include users, a healthcare centre, and a service provider (cloud), the healthcare centre outsources the disease detection service to the service provider (cloud), and the users request health information from the service provider (cloud). In this scenario, we focus on creating a cloud-based multi-party online disease diagnosis system that respects the user's personal information privacy, the healthcare center's data confidentiality, and the system's efficacy.

- In healthcare information relay system models, which consist of users, healthcare center, and gateways, clients conduct health queries from the healthcare center. The query information packs are relayed by the gateways during the query procedure. We allowed the gateways to conduct some preliminary processing in order to make the e-Healthcare more efficient. In this context, we focus on building a medical packets relay protocol for an emergency medical system, while preserving the user's personal information privacy, the healthcare center's data confidentiality, and the systems' eciency.

CHAPTER 2

Literature Survey

2.1 Literaure Review

Pseudonymization, data encryption, access control, and privacy-preserving data outsourcing are some of the strategies used to acquire the secure e-Healthcare system. We'll go through four distinct sorts of privacy-preserving approaches that may be used with an e-Healthcare system in this chapter:

1) Pseudonymization scheme
2) Access control scheme
3)Data encryption scheme
4)Differential privacyscheme

We will also go through some works on illness detection. Following that, we'll go through the works applying these techniques to achieve privacypreserving e-Healthcare network.

2.1.1 Pseudonymization based scheme

The initial method employed for the privacy-preserving e-Healthcare systems was called Pesudonymization. Several pseudonymy research projects have emerged in the previous years. One of the main concepts behind it is to erase any data that might potentially be used to identify people. To maintain anonymity, before discussing or publishing something, one uses a pseudonym. There is no connection between the attacker's alias and the victim. The categories of patient data acquired are data that are useful to the end user, and data that is pseudonymous. The personal information in these pseudonym schemes not only denies any link between the user and their true identity, but it is also stored in a safe way. In a roundabout way, privacy was solved when pseudonyms were created. Despite this, pseudonymization is a poor mechanism for safeguarding privacy when data is gathered from many data sources.

A notion defining a database is K-anonymity.. The concept of quasiidentifiers, or qualities that may be used to link records to people in a database, is introduced first. According to K-anonymity, the quasi-identifiers of every record in a database are linked to no fewer than k users. To achieve k-anonymity, a variety of applications and methodologies are presented. We propose the use of k-anonymity to protect privacy in wireless sensor network medical environments. Another contributor introduces the KMSS (k-member cluster seed selection technique) and shows how it may be used to achieve k-anonymity in k-member clusters. Recently, a proposal presented a scalable k-anonymization approach using Map Reduce for variou medical databases. K-anonymity is required to ensure privacy when transmitting data. If attackers offer information on a single user related to a group of quasi-identifiers, all users related to those quasi-identifiers will be recognised.

2.1.2 Access control scheme

The access control rules are privacy preservation techniques. Hybrid access control policies are used most of the time to provide privacy-preserving access control techniques. It is usual to combine access control with pseudonymization in a single privacy-preserving method that anonymizes users' data and distributes the anonymized data. patients having control over who is given access to their sensitive medical information is the objective of a patient monitoring system (PHI). Patients sign a contract with the healthcare centre detailing their usage of PHI. Role-Based Access Control (RBAC) has long been considered a decent option for implementing security controls, and now a cloud-based solution has been presented to help improve it. Additional features include a larger RB context-based access control set, the ability to communicate with many cloud servers, and the ability to delegate authentication. The purpose, roles, terminology, and structure are all essential to succeeding in this endeavour.

2.1.3 Data encryption scheme

Data encryption is often used in privacy-preserving e-Healthcare systems. The genuine identities of users are encrypted and saved as pseudonyms in some pseudonymization techniques.

Furthermore, encrypted patient data is used in a number of privacy-preserving e-Healthcare systems.

There are two forms of homomorphic encryption: partial homomorphic encryption (PHE), which enables encrypted data to be added, and fully homomorphic encryption (FHE), which enables data to be added and multiplied.

The most widely used homomorphic encryption technologies are ElGammal and aillier. The completely homorphic encryption, in particular, achieves the health query on encrypted data. However, its implementation demands a significant amount of time-consuming calculation. In recent work, a new parallelization to speed up the totally homomorphic based scheme was presented. Figuring out how to make scheme efficient when implementing homomorphic encryption, especially in major health databases, is a major difficulty. Many individuals have lately looked at a privacy-preserving health monitoring plan based on encrypted medical data and wearable gadgets. Other projects improve the privacy of patients' location in mobile medical inquiries.

In the paper, Aydın et al. offer a privacy-preserving data destruction strategy that utilises the prior encryption. The recommended approach is inefficient in making the budget-preserving allocation. KATZENBEISSER et al. offer recommendations for embracing privacy-reserving schemes that use homomorphic secret-key systems such as the Paillier cypher. It can take a long time to run the public key encryption system. Key management becomes a serious issue even in symmetric encryption-based systems, which are commonly efficient. In this thesis, we implement privacy-protecting e-Healthcare rules by using data encryption techniques.

2.1.4 Differential privacy based scheme

Differential privacy is a technique for calculating a measurable privacy constraint depending on the disclosure of a query result. In statistical analysis, it's often used. To discover crucial places on DNA, they used these differential privacy methods. A paper recently proposed a differential privacy scheme for privacy attack on Genomic Beacon Services. Unfortunately, the differential privacy based method isn't ideal in various medical circumstances that need exact outcomes due to the randomization added to the results.

2.1.5 Disease Detection

Since early diagnosis of disease can reduce side effects, safety risks, and the financial burden, it's of interest to medical and bioinformatics experts. Another useful example of rule-based recommendation system development is provided by Nooj et al., who established a rule-based recommendation system for the prediction of heart disease in 2012. A multivariate logistic regression technique was utilised by Bouwmeester et al. in 2013 to build a risk prediction model in which a mix of variables for various symptoms and the environment is utilised to fit a linear transfer function.

Big data analytics and preserving users' and health-related data privacy are all vital topics of research when it comes to finding ways to discover disease risk. This thesis includes a risk prediction model, and a number of privacy-preserving privacy-preserving e-Healthcare systems are offered as a result.

2.2 Fundamental Concepts

2.2.1 Bilinear Pairings

Bilinear pairings have been widely employed in the design of key agreement, signature, and encryption protocols. Allow G and GT to be two multiplicative cyclic groups with about the same prime order q. Suppose G and GT are equipped with a pairing, i.e. a non-degenerated and efficiently computable bilinear map e: G ->G! GT , with e(ga1, gb2) = e(g1, g2)ab 2 GT.The computational Die-Hellman (DH) problem is difficult for any a, b 2 Z-> q, and any g1, g2 2 G in group G. It is impossible to compute gab in a polynomial time given (g, ga, gb) for g 2 G and unknown a, b 2 Z q. The Decisional Die-Hel, on the other hand. In other words, given (g, ga, gb, gc) for g 2 G and unknown a, b, c 2 Z q, it is simple to check e(ga, gb) to see if c = ab mod q. e(gc, g) = e(gc, g) = e(gc, g).

Definition 1: Bilinear parameter generator gen is a probabilistic algorithm which takes a security parameter k as inputs and produces a 5-tuple (q, g, G, GT, e) output, where q is a k-bit prime number, and e: G->G! GT is a non-degenerate and computationally efficient bilinear map.

2.2.2 Bilinear Pairing with Composite Order

Bilinear groups of composite order are those that have an optimum bilinear map in which the group order is a product of two large primes numbers, allowing for homomorphic public key encryption. Set N = pq, with p and q being two huge prime numbers with the same bit length. The cyclic groups G and GT belong to the same composite order N. G and GT are called bilinear maps with composite order e: if there is a computable mapping. G to G! GT wth the following charactertie:

- **Non-degeneracy**
- **Computability**
- **Bilinearity**

The following are the definitions of a composite bilinear generator and the subgroup decision problem:

Definition 2. Gen is a probabilistic algorithm which takes a security parameter as an input, and outputs a 5-tuple (g, N, G, GT, e), where N = pq, p, and q are two bit length distinct prime numbers.

Definition 3. The problem of subgroup decision is shown as follows:

If g be a generator of G, then g1 = g2, G can generate the subgroup Gp = g01, g11,..., gp11 of order p, and g2 = gq 2 G can generate the subgroup Gq = g0 2, g1 2,..., gp1 2 of order q. Determine Gp when given a tuple (e, G, GT, N, h), where h is drawn randomly from either G or the subgroup Gp. The hard subgroup decision problem assures the security of the BGN homomorphic cryptosystem

2.2.3 BGN Cryptosystem (Homomorphic)

Boneh, Goh, and Nissim proposed the BGN cryptosystem, which is the first "Somewhat homomorphic" cryptosystem with a consistent cryptotext. The BGN cryptosystem's key concept is based on the subgroup decision assumption, which supports a polynomial number of additions but only one multiplication. In a nutshell, it has three functions: key generation, encryption, and decryption:

Gen(k): Based on single parameter k, find 2 k-bit prime numbers p, q, and set N = pq. Let g be a generator of G with the order N. Find a mapping that is mappable: G ->G! GT. Set h = gq, which would be a generator of the subgroup G with the order p.

Combine the public and private keys pk = (N, G, GT, e, g, h) and sk = p. enc (pk, m): enc(pk, m): enc(pk We choose a random value r 2 ZN when given a message m from a small space and output the encrypted text = Enc(pk, m) = gmhr 2 G.

Dec (SK, C): Provided a ciphertext of size C and the secret key SK = p, do the calculation p = (gmhr) p = (gp) m. The message m comes from small spaces, so it attempts to solve the discrete log of (gp) m with the base gp.

The following are the proerties of BGN cryptosystem:

• Addition in cryptotext G: Given two cryptotexts Enc (m1) and Enc (m2) 2 G, we will have Enc (m1) • Enc (m2) = Enc (m1 + m2).

• Addition in ciphertext GT: Given two encryption texts EncT (m1) and EncT (m2) 2 GT, we will have EncT (m1) • EncT (m2) = EncT (m1 + m2).

• Multiplication from ciphertext G to GT: Given two cryptographic texts Enc (m1), Enc (m2) 2 G, we have e(Enc(m1), Enc(m2)) = EncT (m1 • m2) 2 GT.

2.2.4 Max heap

The max-heap is a compound binary tree in which the value of each internal node is larger than or equal to the value of that node's descendants. The max-heap is widely used in the top-k ranking algorithm. In particular, as compared to exhaustive element by element ranking, max heap is particularly efficient for identifying the top-k items when the order of these k items is not important. as seen in Fig. 2.1, The elements of a max-heap may be easily mapped into an array: If a node is saved at index k, its left child is stored at index 2k + 1, its right child at index 2k + 2, and its parent, if one exists, is saved at index (k 1)/2. A max-heap may be effectively represented with a simple array since it is a full binary tree. Furthermore, given a set of N values, a heap holding those values may be formed by simply shifting each internal node to its proper location.

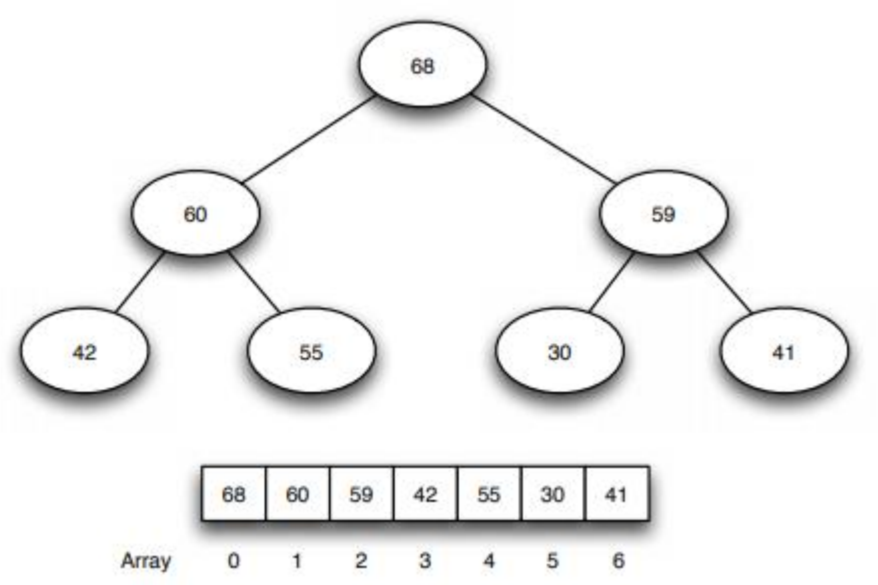


Figure 2.1: A sample of max heap

The inserting algorithm is usually engaged in filtering nodes from the leaf to the root when inserting a node into a max-heap with N nodes. The number of steps required to filter values down will be minimised if the heap is full, which implies N = 2d and the heap height is d. The cost of adding a node is O (d), or O (d) (lg N) depending on the parameter value.

2.2.5 Pseudorandom Function

Pseudorandom Function (PRF) is an efficiently computable function fk (.) whose value is unidentifiable from random components in the function range for a randomly chosen key k.

Definition 4: Algorithms F, on input K 2 {0, 1} x 2 {0, 1}⇤ output {0, 1} F : {0, 1} ⇥ {0,1}⇥ !{0,1}.

Definition 5: For all probabilistic polynomial-time function D, a variable-input-length pseudorandom function F is a variable-input-length pseudorandom function F: r [DF (K,) () = 1] r [DF (K,) () = 1] r [DF (K,) [DF (.) () = 1] r 1, f is a uniform choice of F, where ngl(n) is a negligible function in K 2 , and ngl(n) is a negligible function in K 1.

Definition 6: The oblivious PRF or OPRF is a protocol with an RF F (k, x) in which a sender S sends k to a receiver R via an input x, the R computes the value F (k, x), but the sender S learns hardly anything about the cryptotext.

2.2.6 CCA2 Public-Key Encryption technique

Public-key cryptography is an asymmetric method that uses two keys: a public key and a private key. A public key that may be widely distributed, as well as a private key that only the person who creates the cryptographic system owns. Three algorithms (Gen, Enc, and Dec) make up the publickey encryption technique. Gen is a polynomial-time probablistic key generation method. Dec is a deterministic polynomial-time cryptography algorithm, whereas Enc is a deterministic polynomial-time encryption algorithm.

Definition 7:A two-stage adversary against public-key encryption = (1, 2) We'll start with a basic experiment using this method. The advantage chosen-ciphertext may be attacked through public-key encryption (KE). A KE, we contend, is indistinguishable from adaptive ciphertext attacks (IND2). The advantage of n in the experiment is a negligible function of n if all probabilistic polynomial-time (T) adversaries make a polynomial number of oracle queries.

2.2.7 Model of Disease Risk

Many diagnosis forecast models incorporate variables and environmental data.to predict the severity of a certain disease .The association between various diseases .The odds ratio (OR), which is the ratio of odds in a symptom and symptoms of disease, expresses the stage of a disease.The OR of a certain disease Yi for the symptom predictors Ai = {a1, a2,a3 ··· , am}, with each predictor value aj {0, 1} for j = 1, 2, ··· , m, is generally represented in the terms of regression coefficients Bi = {b1, b2, ··· , bm} of same length m. In other words, the regression coefficients works as the weights of the symptom predictors changes. In this way, the predicted risk of the disease Yi with respect to certain symptom Ai can be computed as:

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

In this model, where is an estimated intercept? This model has been widely employed in the sectors of medicine and healthcenters for disease risk assessments and cure.To simplify the risk score calculation, the overall disease risk score S corresponding to the risk of the diseases = (Yi =1|i) = 11+exp(S) may be computed as follows:

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

2.2.8Disease Risk Threshold Determination

The regression coecients Bi = b1, b2, •••, bm, and the estimated intercept for predicting some disease Yi may be obtained from the logistic regression model using a large volume of real-world medical data. We can create a disease risk threshold S to assess if a user UI with the symptom predictors I = (a1, a2, •••, am) has the disease Yi with a probability. If +m j=1 aj •bj S Then, we may deduce that user Ul has a high probability of having the disease Yi. Otherwise, when + Pmj=1 aj · bj < S th,we infer that Ul has Yi with a low probability. Since this disease risk model is an asset, the values (Bi = b1, b2, •••, bm,, S th) should be done privately (i.e., the rivacy-reserving requirement). In the next section, we will demonstrate our Guide scheme, which implements the disease risk model in an efficient and privacy-preserving manner to get pre-clinical guidance for medical patients.

CHAPTER 3

Problem Description

3.1 Introduction

Commercial integration of processing clinical analytics has progressed while maintaining the privacy of personal medical data. How can we be sure that our personal health information hasn't been made public and exploited by third parties such as insurance companies or future employers, for example? Data leaking is a serious worry for service providers, in addition to a patient's desire for privacy. Medical practitioners in nations like the United States are subject to stringent rules. A regime is the Health Insurance Portability and Accountability Act (HIPAA).

We propose an efficient privacy-preserving pre-clinical guidance scheme to solve the above mentioned privacy challenges and improve the accuracy of disease threat prediction (and resulting in more accurate disease threat predictions). Using the guidance scheme provided in this chapter, users may undertake pre-clinical diagnosis based on their health profiles and acquire suggestions from credible sources (e.g. hospitals and medical service providers). Furthermore, the data was supplied to hospitals and other medical service providers in order for them to quantify illness risk using illness predicting models in a secretive manner. After that, we show how our suggested Guide system protects both the individual customer and the medical service provider's privacy. We show that our suggested Guide method is cost-effective in terms of computation and transmission across vast distances based on the results of our experiments.

3.2 Solution

We look at a distributed system where data is propagated in a diverse way. This means that different sources gather different sorts of data for the same collection of data. In contrast, with homogenous data distribution, a variety of sources collect the same bits of information on a variety of topics. It is easier to deal with the second instance because there is no genuine integration to be done. As a consequence, we solely take into account the first scenario. Each source Si has a connection Ri with parameters (ai1, ai2...., aik), where ai1 uniquely specifies tuples in the connection Ri, suggesting that ai1 is the connection Ri's unique identifier. We presume that no other property or set of attributes, other than ai1, can be used as a candidate key for the relation Ri without suffering genericity loss. We assume that the main key for each relation is the same, and that the join of two or more relations is always computed using the primary key value, i.e., only those tuples in the relations RI and RJ for which ai1 = aj1 may be joined. If ai1 = aj1, we argue that the relation with attributes (f (ai1), ai2,..., aik aj2,..., ajm) is a privacy preserving join of relations Ri and Rj, but it is operationally infeasible for any party, including data. Thus, in this project, we are tackling the problem of computing the privacy preserving join of n relations; where source Si owns a relation RI (1 I n).

In DaaS, twofactors Exacerbate the problem of privacy.

To begin with, DaaS services acquire and keep a large quantity of personal data about its consumers. Second, DaaS services offer the capability of sharing this data with other businesses. Furthermore, the development of analytical tools has made it simpler to analyse vast amounts of data for study, raising the risk of security breaches. We'll use our epidemiological scenario to demonstrate the privacy issues that arise during service composition.

Privacy Specification:

Both the input and output variables (SSN and DNA) are considered sensitive data by scientists. Consider the following idea put out by this scientist: “weatherconditions” has an influence on the H1N1 virus.” The scientist may prefer to keep invocation private (regardless of what S5.1 captures and provides as data), since this might give competitors access to sensitive information. The first problem outlined above highlights the need for a formal model that defines what private data is and how it is defined.

Privacy within compositions:

Important services (which participate inside a composition) may need the submission of data that other services cannot disclose for privacy reasons. They may also have conflicting security concerns regarding the information they've shared. Assume S1.1 states that it will share its data (SSN) with a third-party service for a limited time. S3.1, on the other hand, ensures that acquired data (SSN) is used for an infinite amount of time. S1.1 and S3.1 have different privacy constraints when it comes to the SSN. This will invalidate the issue in terms of privacy issues.

Dealing with incompatibilityin privacy:

There are policies in compositions. The mediator's mission is to find similar replacement for outmoded services while keeping confidentiality. When dealing with instances involving contradictory privacy policies, the easiest solution is to just refuse them. A more fascinating, if challenging, strategy would be to try to create a consensus among constituent services to resolve their privacy incompatibilities by expanding the number of composition plans given by the mediator.

CHAPTER 4

Proposed Solution

We discussed the privacy-preserving healthcare query in a two-party system model in the previous chapter, which only includes the users and the healthcare service provider. In this chapter, we'll look at a three-part system model in which the healthcare center outsources computation tasks to the cloud server, and users access the cloud server for healthcare information.

4.1 Introduction

Body Sensor Networks sprang from research on wireless sensor network technologies and applications a few years ago (BSNs). Health inquiry services have gained a lot of attention, and they're getting more popular, thanks to the growing prevalence of smart phones and BSNs. A user is furnished with a wearable BSN that contains wireless physiological sensors for motion detecting, such as smart textiles, skin electrodes, surface thermistor, or three-axis gyroscope. In recent years, other BSN applications have been proposed, including early detection, elder support, physical activity tracking, and so on. Smart gadgets, such as smart phones and smart watches, are used to capture and analyse these sensitive data streams. One of the benefits of the wearable health system, as illustrated in Fig.4.1, is the ability to utilise the health tracker service without interfering with the user's capacity to complete daily tasks.

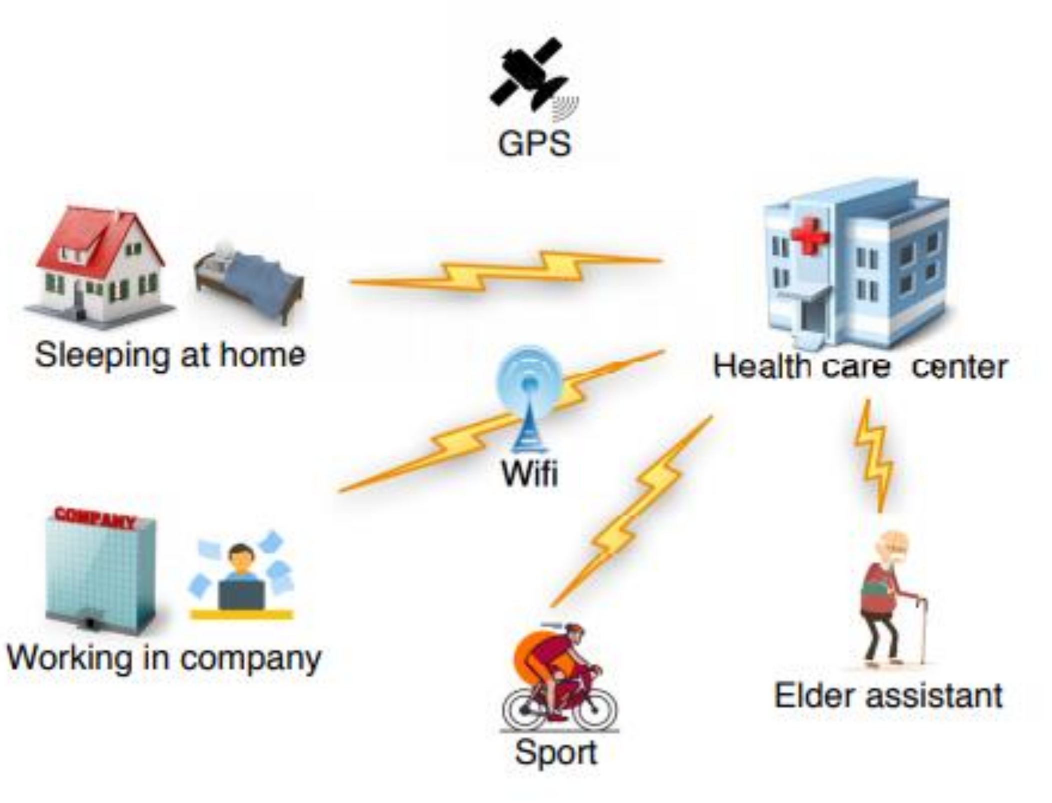


Figure 4.1: Wearable health monitor system

While the BSN medical query service is well-known and useful, most health inquiry services need users to provide physiological data, medical history, and personal information, which raises worries about data breach and abuse. The health service provider prefers to outsource a large amount of health data and health query services to the cloud, which can execute high-performance processing, because of the operational and financial benefits. As a result, protecting users' privacy in health Query has piqued interest. Health data and diagnosis models, on the other hand, are intellectual property of the health centre and should be kept apart from the crowd. The obstacles of designing a privacy-preserving health inquiry on an externally sourced cloud, which sits at the convergence of mobile computing and artificial intelligence, are outlined below:

• **Challenges for health query on encrypted data**:

The health-related inquiry service Because both the provider and the user are unwilling to reveal private health and psychological data, both the health and psychological data are encrypted before being sent to the cloud in an outsourced health inquiry service. It's tough to do a health inquiry on encrypted data. Rare privacy-reserving procedures, such as differential privacy, may stort the data, making it unfit for medical study and, in some circumstances, deadly to the users. As a result, for medical calculation and analysis, the health query scheme on encoded data should be accurate.

• **Challenges forefficiency on health query and encrypted data**:

For executing health inquiries and data analysis on encrypted data, many cryptographictechniques have been created. For example, many homomorphic encryption techniques have been described. The overhead of the calculation, on the other hand, becomes a stumbling block, preventing them from becoming widely used at a lower industrial level. In order to deliver acceptable user experiences, the health query scheme on encrypted data should be efficient.

• **Challenge on security of data**:

Despite the fact that health data and users' psychological data are encrypted, it is plausible to presume that an intruder has access to the plaintext of some personal information or other side data. As a result, requiring the scheme to be robust against known-plaintext attack is essential. In order to prevent illegal access, only authorised users are allowed to perform health queries.

In this chapter, we present HeOC, a health query scheme over outsourced cloud that allows authorised users to query sensitive health information without providing their Sensitive physiological data while maintaining confidentiality and integrity. We create a unique anomaly detection methodology that effectively detects anomaly disease with a high likelihood of requiring further study. The key contributions of this chapter are listed below:

- **Firstly,**The HeOC technique is proposed as a way to deliver an efficient and privacy-preserving health inquiry using an outsourced cloud. The user initially detects the anomaly disease with the sensor collected data in order to sort out the suspicious disease as well as its related physiological data in this scheme. The user may then utilise a number of OPRF protocols to query the precise illness level using the filtered result from the sensor anomaly detection approach. Experiments show that our HeOC strategy is effective.

- **Secondly,**SADS is a new sensory anomaly detection methodology for detecting high-risk illnesses that we propose. In the SADS technique, the health care provider uses an encrypted health graph to refer to the crowd. Authenticated users provide encrypted physiological data to the cloud in order to identify high-risk illnesses without disclosing their personal health information.

- **Thirdly,**To evaluate the HeOC scheme's performance in a real-world situation, we construct an Android application and two Python programmes. The findings reveal that the HeOC system is effective in terms of communication and computation. The rest of this chapter is laid out as follows: In Section 4.2, we describe our system model, security need, and design purpose.

4.2 Design Goaland Models

We formalise the system model, security requirements, and define our design goal in this section.

4.2.1 System Models

Two distinct settings for the rivacy-reserving health query have been investigated: two-party health query and outsourced health query. This chapter focuses on the outsourced health inquiry context. As indicated in Figure 4.2, the system is made up of three types of organisations: a health care provider, authorised users, and the cloud.

• Services in the medical field The provider has a lot of health information. Hospitals or professional medical organizations are examples. We'll utilise the hospital as the owner of health data to keep the system model simple. The hospital employs a variety of diagnostic models for a range of disorders, and these diagnostic models are used to order health-care services. Because of the economical and operational benefits of data source, the hospital requests the health inquiry via the cloud. The hospital, on the other hand, is adamant about not disclosing the valuable diagnostic models to the general public. The hospital then encrypts the diagnosis model and health data before sending the encrypted data to the cloud.

• The hospital validates the identification of the users. Each authenticated user receives a wearable BSN, which consists of a number of wearable sensor nodes that remotely record and analyse physiological data.

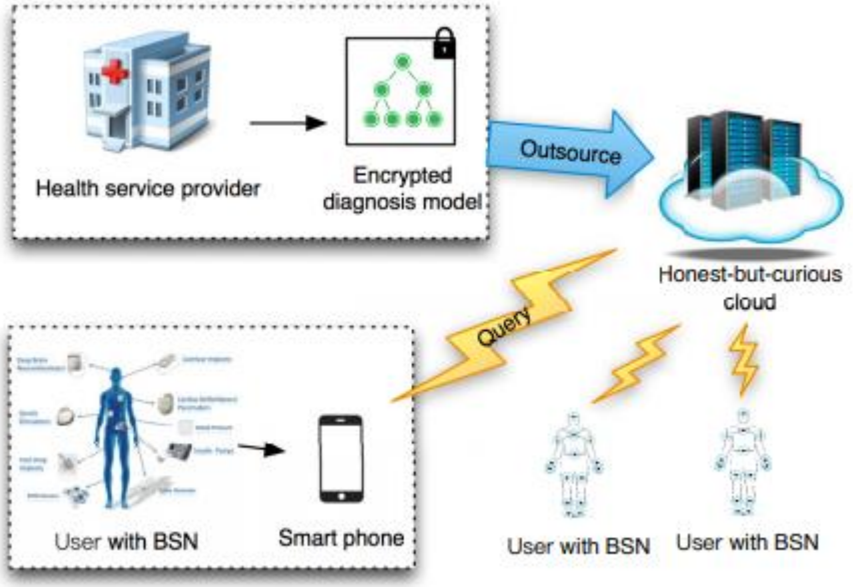


Figure 4.2: System model of outsourced health query

A smartphone phone is used to process the data obtained. To avoid exposing a user's privacy to the cloud, the user acquires an access key in advance, encrypts the query, and sends it to the cloud using cryptographic and privacy-preserving mechanisms.

Clients submit encrypted queries to the cloud, which utilises the hospital's encrypted diagnostic model to perform illness diagnosis utilising cryptographic and privacy-preserving techniques. The cloud then sends the encrypted diagnosis result to the users.

4.2.2 Security and its Requirement

The cloud is semi-honest in the arena of outsourced structural model; in this case, the reputation for correctness of the outsourced diagnosis service depends on the correctness of the underlying model. The cloth may also be placed in In order to maintain the secrecy of the diagnostic model and maintain the privacy of the patient's medical data, the following security requirements must be met:

- **Confidentiality.**The diagnosis models of the hospital should not be disseminated to the public, since these models represent the hospital's intellectual property. As previously mentioned, the cloud should never be able to recover the diagnostic model using encrypted medical data obtained from the hospital, implying that the diagnostic models should be secure under a text-only attack. Furthermore, even if a attacker understands the plaintext of a certain encrypted diagnosis model, it should not be able to get the plaintext of any other encrypted data in the proposed method.

- **Privacy**.Each member of the HeOC system executes health-related inquiries, such as age, gender, and medical history, using sensitively gathered physiologically and personally data. As opposed to a patient's PHI, the cloud should not have access to patient-specific PHI (such as medical records) by looking at the ciphertext of a query data value, suggesting that the HeOC scheme should be secure solely under ciphertext attack. If attackers have the plaintext for encrypted profile data for which they don't have the corresponding encryption key, they should never be able to decrypt the user's physical data that is contained in other ciphertexts. That in other words, under the HeOC approach, the system is secure.

- **Authentication.** Only consumers who have been specifically authorised are permitted to ask the cloud for diagnoses and treatments linked to their health. Illegal users that send health queries should be traced as soon as possible. Additional security controls, such as access-pattern and role-based access control, may be possible. Only the security standards specified above are evaluated in this chapter. There will be no more inquiry left to do on the remaining items.

4.2.3 Goal of Design

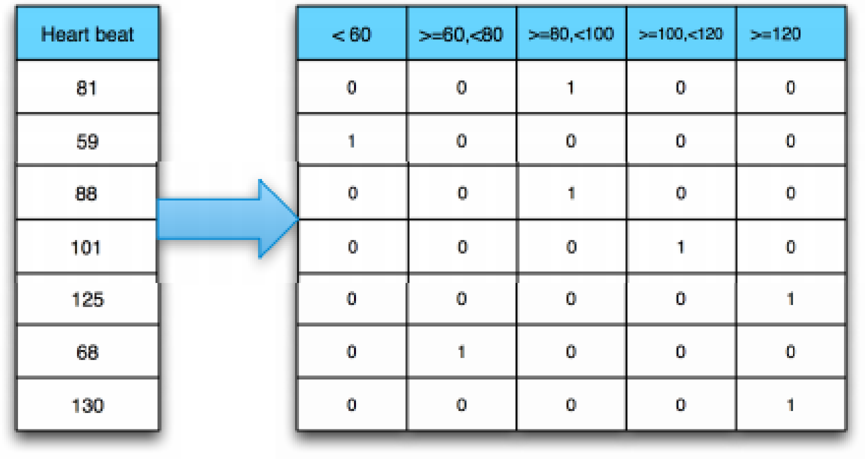
To obtain an efficient and privacy-preserving health inquiry, we'll develop a system using outsourced cloud computing and based on the model discussed previously. To wit, the following three objectives should be met in particular:

- **Accuracy Goal.**to verify the health investigation's accuracy Even though privacy-enhancing measures can't guarantee precise diagnosis, because inconsistencies might have significant impact on customers, they should nonetheless be used to avoid the risk of inaccuracy. Due to the simplicity of the proposal, the system must be highly accurate.

- **Security Goal.**Attack techniques that use just ciphertext and attacks that use only known plaintext should be resistant to the suggested approaches. in order to maintain the anonymity of the diagnosis model, it is imperative that the user's physiologic data, as well as the diagnosis model, remain secret.
- **Efficiency Goal.**Processing complexity should be avoided because of the need for real-time health investigation services and the variety of users. Examples of exponential returns, for example, should not be based on the formula.

4.3 HeOC Scheme

We propose a framework for privacy-preserving inquiry based on a sensor anomaly detection approach, pseudodorandom function RFs, and symmetric encryption methods in this section.



Heart beat
81
59
88
101
125
68
130

< 60	>=60,<80	>=80,<100	>=100,<120	>=120
0	0	1	0	0
1	0	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1
0	1	0	0	0
0	0	0	0	1

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

4.3.1 Brief Overview

The suggested HeOC approach may identify sensory abnormalities while simultaneously evaluating sickness severity while keeping the privacy of the patient intact. Wearable BSNs (or biosensors) are provided to all members of HeOC, and these BSNs remotely acquire and interpret physiological data. The user's smartphone gets hold of the raw data and processes it further. Any information that pertains to the user, such as age, gender, medical history, and sensory data, is considered personal information. In order to begin our proposed method, we first need to briefly provide the facts about these users' particularities.

The act of converting continuous data into discrete components is known as discretization. With physiological data, we refer to standardized personal information. A set of study parameters is represented by X = x1... xn. Each x 2 X component is standardised from the original personal data. As illustrated in Fig.4.3, x is 4 (00100) for a heart rate of 80-100; x is 8 (01000) for a heart rate of 60-80; x is 16 (10000) for a heart rate of 60-80; x is 2 (00010) for a heart rate of 100-120; x is 1 for a heart rate of 120 or more (00001). Preprocessing makes medical data suitable for numerical and parameter evaluation, allowing them to standardise the algorithm's output, so simplifying the health inquiry.

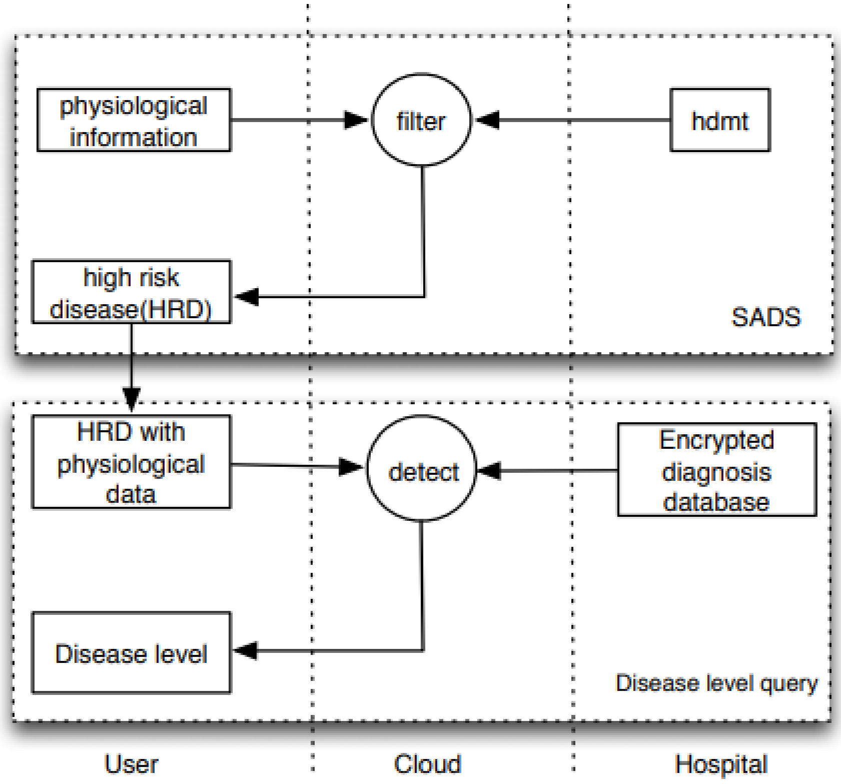
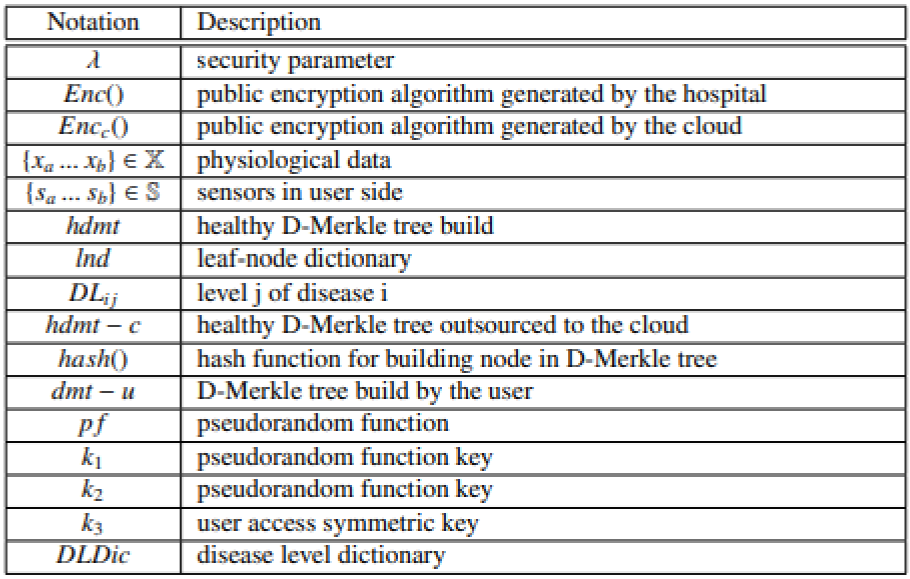


Figure 4.4: Query flow in HeOC

The He's concept combines a sensing and diagnosis technique known as the Sensory Analyzer Detection System (SADS) and a disease level query technique. The query flow seen in Figure 4.4 is shown. With the SADS methodology, a trusted client may detect sensitive data and examine infections with a high risk, all while keeping information private. After recording the physiological data relevant to the high-risk condition, the user next utilises the cloud to search for the exact sickness level in a privacy-protected manner. You'll be glad to know that we've highlighted the most important details in Table 4.1 for your convenience.

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
$\{x_a \ldots x_b\} \in \mathbb{X}$	physiological data
$\{s_a \ldots s_b\} \in \mathbb{S}$	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 D-Merkle Tree

Before we go into the details of the HeOC scheme, let's have a look at the D-Merkle tree, which serves as a foundation for it. A Merkle tree version, the suggested D-Merkle tree is a Merkle tree version. As illustrated in Fig.4.5, the D-Merkle tree exhibits the following features:

1)The crypto hash code of its child nodes' labels is used to label every nonleaf node.

2) Each leaf node represents a disease by pointing to the related health data xa... xb2 X, which are gathered from the sensors sa... sb.

3) Each leaf node has a accompany summary; for example, the accompany description of leaf node D1 in Fig.4.5 is [D2, n2]. Given one leaf node and its accompany list, the root of the D-Merkle tree may be readily rediscovered. In the framework of the user-cloud-hospital disease level detection system, the hospital generates.

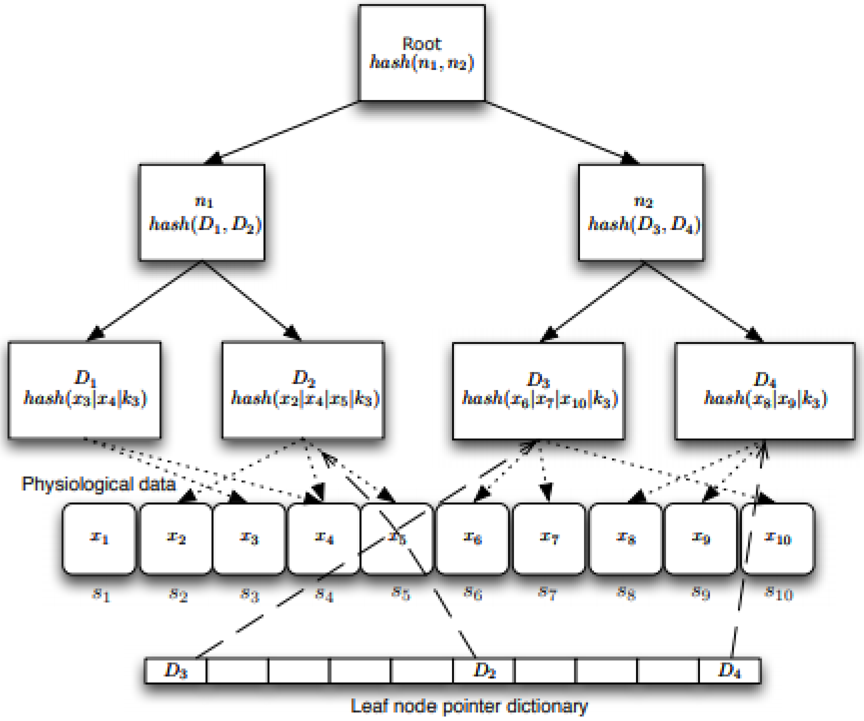
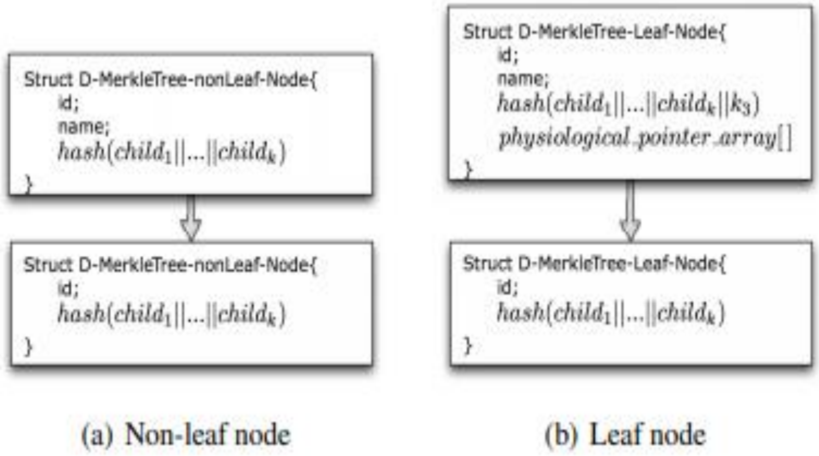


Figure 4.5: An example of Disease Merkle Tree with physiological data $x_1, ...x_{10}$, which are collected from sensors $s_1, ...x_{10}$.

The D-Merkle graph is directed to clients that have been validated, and is open to the public. Due to the need for secrecy, the D-Merkle tree was developed mostly by Hotmail, as shown in Fig. 4.6. D3 is the sister node of n1 if the D-Merkle tree in Fig. 4.5 has just three diseases, namely D1, D2, and D3. The hospital also publishes a leaf dictionary, which uses the nodes (nodeId, leaf node) as (key, value) pairs. It costs $ to find a disease leaf node on this leaf diagonal (1). The risk of identifying a sick leaf node is significant in the He system. The programme will run faster due to this data format.


Struct D-MerkleTree-nonLeaf-Node{
id;
name;
$hash(child_1||...||child_k)$
}
Struct D-MerkleTree-nonLeaf-Node{
id;
$hash(child_1||...||child_k)$
}
Struct D-MerkleTree-Leaf-Node{
id;
name;
$hash(child_1||...||child_k||k_3)$
$physiological\_pointer\_array[]$
}
Struct D-MerkleTree-Leaf-Node{
id;
$hash(child_1||...||child_k)$
}

(a) Non-leaf node (b) Leaf node

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

4.3.3 Setting up the system

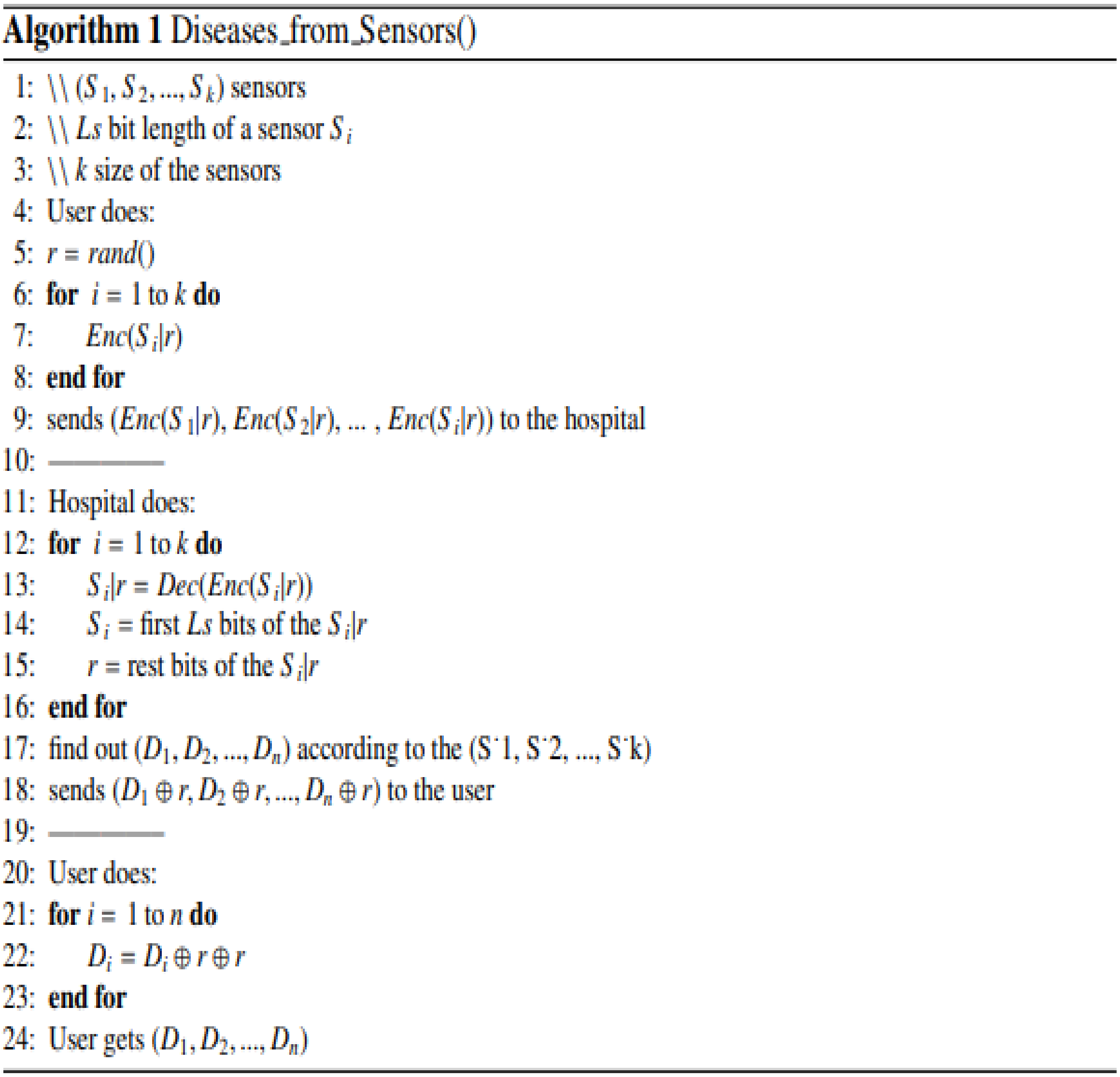
First, the hospital generates an IND 2secure public key encryption algorithm, then releases the key pair and encryption algorithmEnc (). The cloud also develops an IND-2 safe public key encryption algorithm and shares both the public key and the encryption method Encc (). For many disorders, the hospital then generates a diagnostic table based on a multi-level diagnostic paradigm. The disease level DLi j, as well as the related physiologicaldata xa... xb, are listed in each row of the diagnostic table. xa... xb physiological data As noted in section 4.3.1, 2 X are standardised, and each value x 2 X, x MaxLevel.

Plants a healthy D-Merkle tree in the hospital. A D-Merkle tree and a leaf-node dictionary are constructed at the hospital by deciding on the healthy number of all the conditions in the diagnostic table. The node IDs are used as aliases for each sensor ANOMALY DETECTION approach's leaf nodes. The total cost of searching for the HDMT is log(d). cost of applying the leaf node disease level DLi classification LND to identify the leaf node for disease level DLi (1). A short delay occurs before the D-Merkle tree (or subtree) is uploaded to the cloud, after which it is processed for secrecy. Since there are no more illness names or physiologic pointer arrays in the cloud, the leaf nodes of the D-Merkle tree there are empty of disease names and physiologic pointer arrays. This tree is not for us. The id used by the cloud to differentiate leaf nodes is utilised to identify nodes when conducting the test algorithm later. Fig.4.6 illustrates the process for making the hdmt. in this fashion, k3 is a symmetrical number.

Registering users

This suggested method uses an example of the user registration algorithm to demonstrate it.
Individuals initially gather their senses data before transferring the info to the hospital. It is decided which diseases will be found by the sensor based on the list provided by the user, and the ailment ids connected with those diseases are then obtained from the leaf node of the hdmt. Using these illness identifiers, the hospital creates a disease identifier list and sends it to the user. Imagine for example, as illustrated in Fig. 4.5, the hospital owns the HDMT and receives the sensor list.
Additionally, the user is taught of the relationship between each condition ID and their corresponding responder sets. In the sensor alias detector algorithm, the hash value of the HDM T's rotary knob is also communicated to the user for authentication in the future. The user also receives

the symmetric key k3 and the hash function hash() through a secure network.



Algorithm 1 Diseases_from_Sensors()

```
\\ $(S_1, S_2, ..., S_k)$ sensors
\\ $Ls$ bit length of a sensor $S_i$
\\ $k$ size of the sensors
User does:
$r = rand()$
for $i = 1$ to $k$ do
    $Enc(S_i|r)$
end for
sends $(Enc(S_1|r), Enc(S_2|r), ..., Enc(S_i|r))$ to the hospital
————
Hospital does:
for $i = 1$ to $k$ do
    $S_i|r = Dec(Enc(S_i|r))$
    $S_i$ = first $Ls$ bits of the $S_i|r$
    $r$ = rest bits of the $S_i|r$
end for
find out $(D_1, D_2, ..., D_n)$ according to the $(S˙1, S˙2, ..., S˙k)$
sends $(D_1 \oplus r, D_2 \oplus r, ..., D_n \oplus r)$ to the user
————
User does:
for $i = 1$ to $n$ do
    $D_i = D_i \oplus r \oplus r$
end for
User gets $(D_1, D_2, ..., D_n)$
```

The Hospital Constructs an Encrypted Diagnosis Database

By building an encrypted database using the diagnostic table, the hospital has met the privacy-protecting illness level inquiry. Security screening functions as a checkpoint in the hospital. The encryption we use is extremely efficient; it is one of the most contemporary symmetric encryptions that can be searched. A variable-input-length pseudorandom function (RF) on the input keys K 2 0, 1 and x 2 0, 1 produces a string that is between 0 and 1. With regard to RF I f, the Hospital has to use two PRF keys, one of which is identified as k1, and creates the encrypted sickness level index. DLi j is encrypted as follows:

T1=p f (k1, xa||...||xb)

T2=p f (k2, xa||...||xb)

X related physiologic data (x..., xb) are concomitant and result in the RF input (x...||...||xb) in the equation above. T1 is used as the key and T2 is used as the value in the dictionary DLDic, both of which are stored in the database. Once the hospital has distributed the DLDic distribution bundle to the cloud, everyone receives DLDic. Healthy D-Merkle tree is encrypted and sent to the cloud prior to any processing. The system stores a list of disease IDs, depending on the user's sensory data, such as, the HDMT's unique identifier, the hashes of the mouthrot and HDMT, the registration key, the secure parameter, two PRF keys, and the PRF.

4.3.4 Sensitive Sensor Detection Technique

SADS is our approach for creating HeOC, which serves as the foundation for our privacy-preserving health inquiry system that utilises the cloud.

Monitoring based Algorithm

The BSN, which may be implanted in a variety of sites on the human body, wirelessly gathers and interprets biological signals, synthesises responses, and delivers the information to the user's smartphone. The client enters information such as his or her gender, age, and medical history, as well as the sensory data gathered. These data are standardized in the smartphone as physiological data x1...xn. For a disease id and the physiological data x1, x2,..., xk gathered by its associated sensors s1, s2,..., sk, the user calculates hashi = hash(x1|x2|...|xk|k3). The user calculates the hash vector in the monitor algorithm for the disease id vector Id = (id1, id2.., idn) gotten from the hospital:

Hash = (hash1,hash2… hashn-1, hashn)

Verification Algorithm

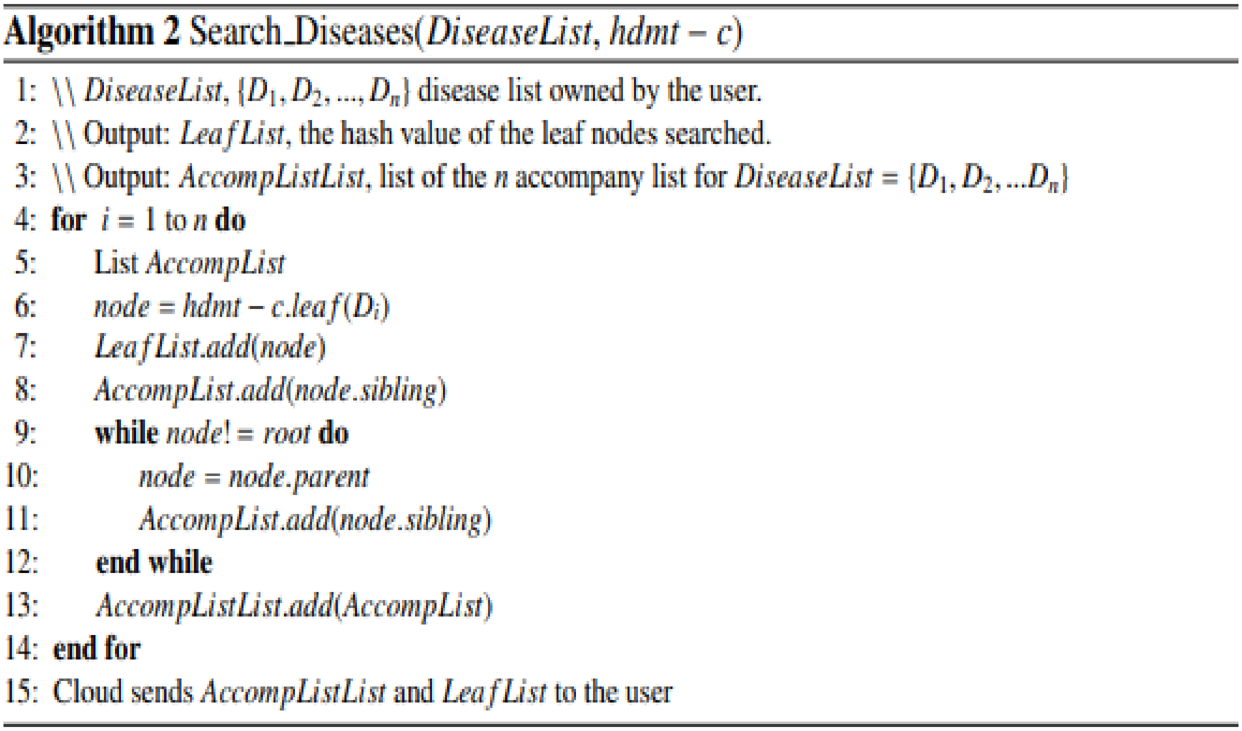
A user must first retrieve disease information from the cloud before doing the comparison. Use the hash value from the HDMT's ROM to validate downloaded information. In order to accommodate a variety of sensors, the data stored in the cloud is different for each user. The method to access the S cloud's sickness data will make sure that the data being used in the future is up to date.

In algorithm 2, each registered user gets the respective leaf nodes of the HDMTC's Tree (Tree List) tied to the sensors they wear, based on how useful the different sensors are.

In addition, the user gets each branch node's unique complement checklist so they may verify the authenticity of each node.

When an algorithm allocates leave nodes, Accomology holds the accounts. At the end of this procedure, the CLOUD restores the user's Leaf List and Accumulator List. Once they have leaf nodes and an enough inventory list, the user compares and validates the information on the user side, which is shown in algorithm 3. To help ensure that all downloaded leave nodes are actually correct, the user computes the hash code of each leaf node and the nodes in the procedure's associated list.

The calculated hash value should match the hash value of the hard drive's read/write head. value with that of the competitors' in order to find which brand offers the best value. It may be represented as: 2 hash = (hash1, hash2, ..., hashn) which was mined from the Momitor algorithm to the counterpart hash value that is added with the new-determined leaf node no-doubt. The patient is at a greater risk of getting the disease if these two levels are out of sync. This results in the AnomalyDiseaseList obtaining Di's disease information, where it will be researched further in the future. Using the verification process, the user goes through the downloaded leaf nodes and filters out the high-risk diseases, which are documented in the AnalysisDiseaseList.



Algorithm 2 Search_Diseases($DiseaseList$, $hdmt - c$)

```
\\ $DiseaseList$, $\{D_1, D_2, ..., D_n\}$ disease list owned by the user.
\\ Output: $LeafList$, the hash value of the leaf nodes searched.
\\ Output: $AccompListList$, list of the $n$ accompany list for $DiseaseList = \{D_1, D_2, ...D_n\}$
for $i = 1$ to $n$ do
    List $AccompList$
    $node = hdmt - c.leaf(D_i)$
    $LeafList.add(node)$
    $AccompList.add(node.sibling)$
    while $node! = root$ do
        $node = node.parent$
        $AccompList.add(node.sibling)$
    end while
    $AccompListList.add(AccompList)$
end for
Cloud sends $AccompListList$ and $LeafList$ to the user
```

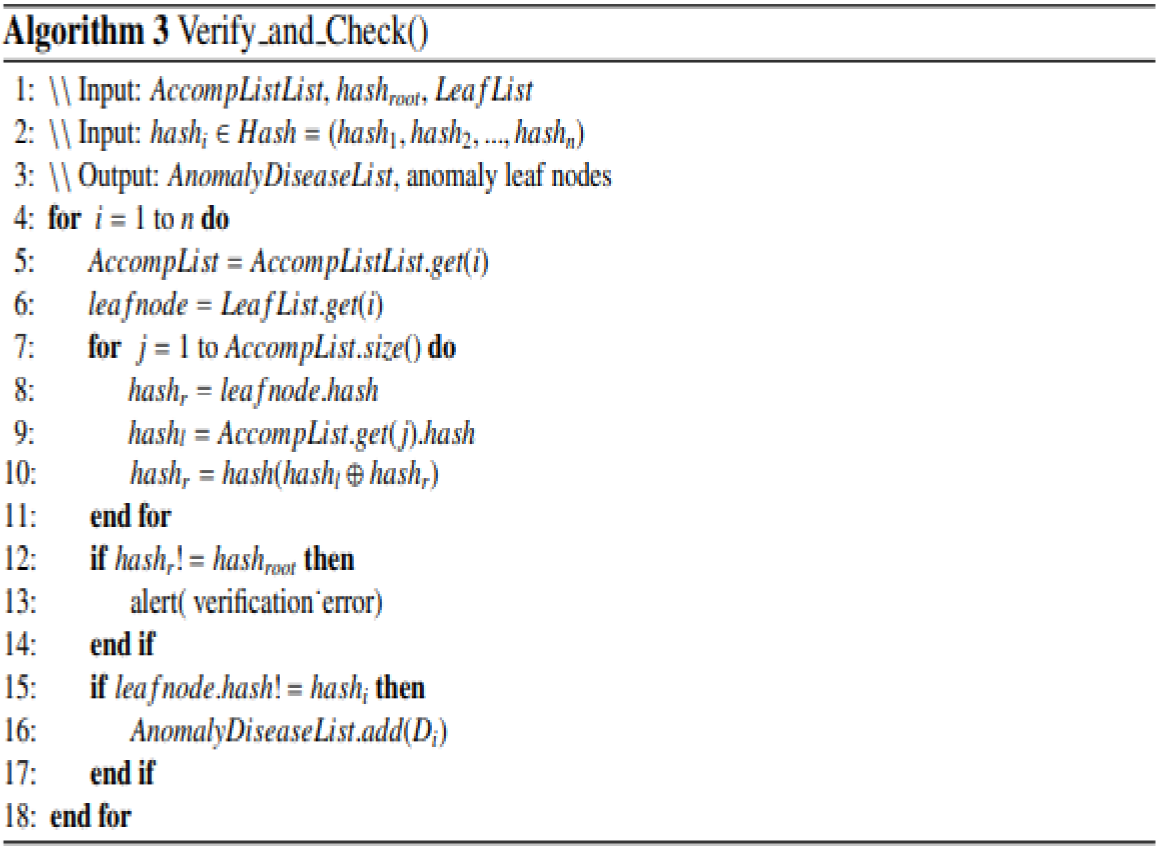
4.3.5 Privacy-preserving Query for diseases

The authenticated client obtains the keys k1, k2 from the hospital over a secure channel. The user uses the SADS approach to locate the anomaly leaf nodes, each of which represents a different disease. We consider one leaf node and one corresponding disease to improve the explanation of the disease category query technique. For the others, nomally leaf nodes, the inquiry technique is the same. The user receives the related (xa... xb) from the physiological point array in the leaf node for anomally leaf node as well as the corresponsing disease. The user calculates two values as follows with pseudonrandom function p f: $t1=p\ f\,(k1,\ xa\|...\|xb)$

$t2=p\ f\,(k2,\ xa\|...\|xb)$

It generates random value r and encrypted t1 using the Cloud's public key cryptography algorithm, Encc (t1|r). The user subsequently transmits Encc (t1|r) to the cloud. The cloud decrypts Encc (t1|r) and receives t1 and r as follows:

$t1|r = Decc\ (Encc\ (t1|r))$



```
Algorithm 3 Verify_and_Check()
\\ Input: $AccompListList$, $hash_{root}$, $LeafList$
\\ Input: $hash_i \in Hash = (hash_1, hash_2, ..., hash_n)$
\\ Output: $AnomalyDiseaseList$, anomaly leaf nodes
for $i = 1$ to $n$ do
    $AccompList = AccompListList.get(i)$
    $leafnode = LeafList.get(i)$
    for $j = 1$ to $AccompList.size()$ do
        $hash_r = leafnode.hash$
        $hash_l = AccompList.get(j).hash$
        $hash_r = hash(hash_l \oplus hash_r)$
    end for
    if $hash_r != hash_{root}$ then
        alert( verification˙error)
    end if
    if $leafnode.hash != hash_i$ then
        $AnomalyDiseaseList.add(D_i)$
    end if
end for
```

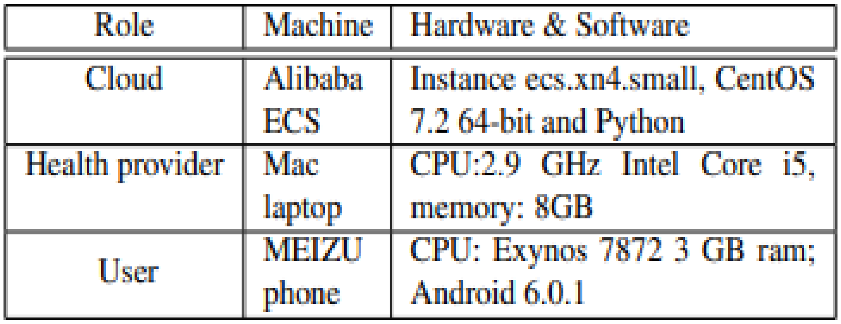
4.4 Performance Efficiancy

We evaluate the performance of the proposed HeOC scheme in terms of computational cost as well as communication cost in this section.

4.4.1 Implementation Settings

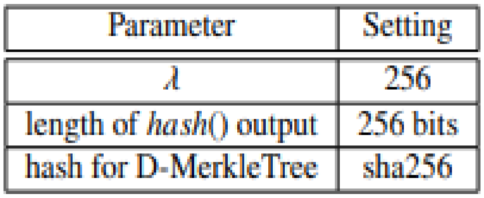
n Python and Java, the HeOC scheme has been implemented. We carried out an experiment using HeOC by using Alibabacom's Elastix compute service, a Mac OSX workstation, and one Android phone in a lab setting. Cloud, hospital, and mobile user all adopt the form of these machines. Table 4.2 lists the soft and hard materials in these machines, whereas Table 4.3 lists the variables. We utilised the Goldreich-Goldwasser-Mali and SH256 algorithms to create RFs.

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 *hash*() output	256 bits
hash for D-MerkleTree	sha256

Hospital staff also has to set up the entire system and complete the encryption of the diagnostic database as part of the system configuration. 12,421 illness types, according to WHO, are recognised in the most recent version of the International Classification of Diseases (ICD-10). Once we determined the time required for SADS, we tested it, as shown in Fig. 4.7. Setup time is reasonable because the system is only installed once.

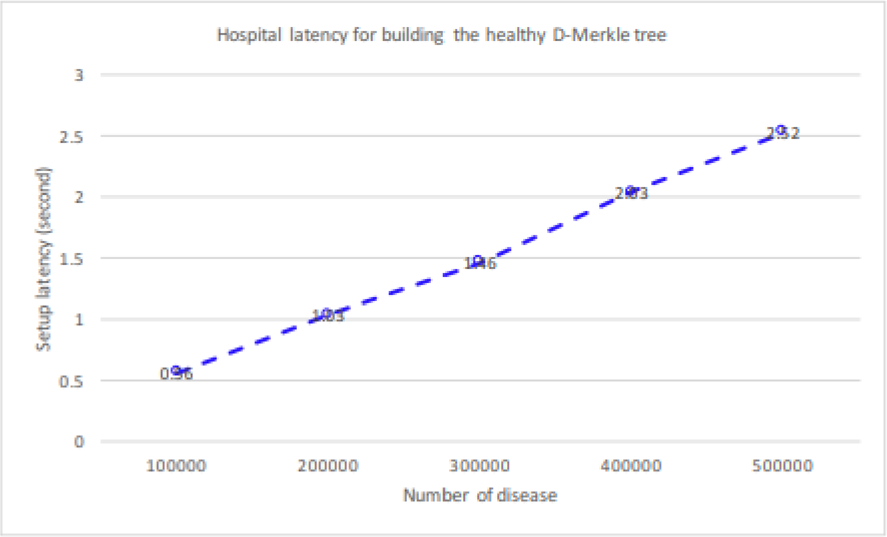


Figure 4.7: Hospital latency for building the *hdmt*.

The user is responsible for three tasks:

1) In the disease level query technique, the user utilises PRF to determine the exact disease level DLi j for each disease with a high risk

2) In the user registration, the user discovers the disease list that corresponds to the user's sensors.

3) The customer downloads the leaf nodes and related accompany list according to the disease catalogue, then verifies and checks the downloaded information in the SADS technique to discover the disease with a significant risk.

To estimate user registration latency, we created 1000 sensors on an Android phone and had it generate 20, 30, 40, 50, and 60 sensor levels between 1 and 1000. In Figure 4.8, you can see the outcomes of the experiments. The setup, which has over 1000 sensors, can be completed in less than 6.8 seconds, even with the current hardware arrangement. The computational cost is tolerable because a user registration is only performed once for each user.

L e e lergreen for verification of the long-term usefulness of the SADS is critical at this stage of the assessment process. It is impossible to accurately determine the verification latency of each sickness without knowing the extent of the high-density (HD) metabolite recorded in the cloud. As a result, we evaluate the verification endurance using the diseases, which are listed as diseases 101, 102, 103, 105, and 107 in the public. You can see the outcome in Fig. 4.9. The verification lagging time increases in relation to the tree's height. At the root level, the HDMT's tree level is 22, and there are 106 diseases in the HDMT. To search for a single sickness, the verification latency is 3 milliseconds. Since all the user needs to do is study all of the information for around one second, users who have a heavy burden of disease to evaluate just have to verify and review it for that time period. It frees up a lot of time for you. For the client, there are n ailments to test, and the HDMT has an L-level tree (nL). The user must provide two prescriptions to query the ailment level DLi j. It was estimated that encrypting an anonymous leaf (which includes such data as GPS coordinates, browser type, operating system, etc.) on PFS will take less than 27 milliseconds, which is imperceptible to the typical smartphone user.

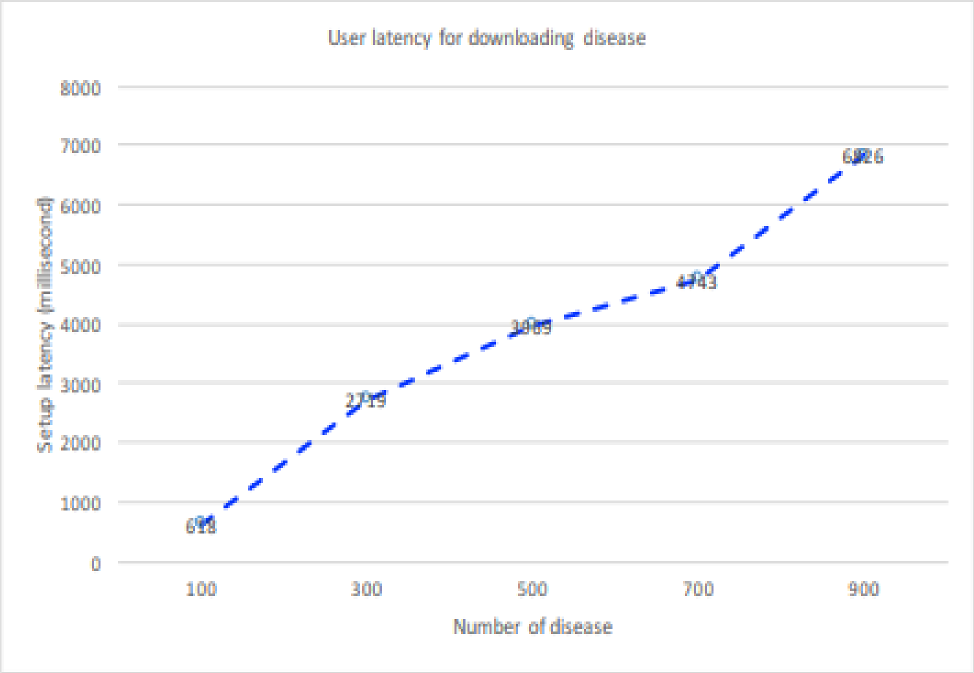


Figure 4.8: User latency for downloading diseases.

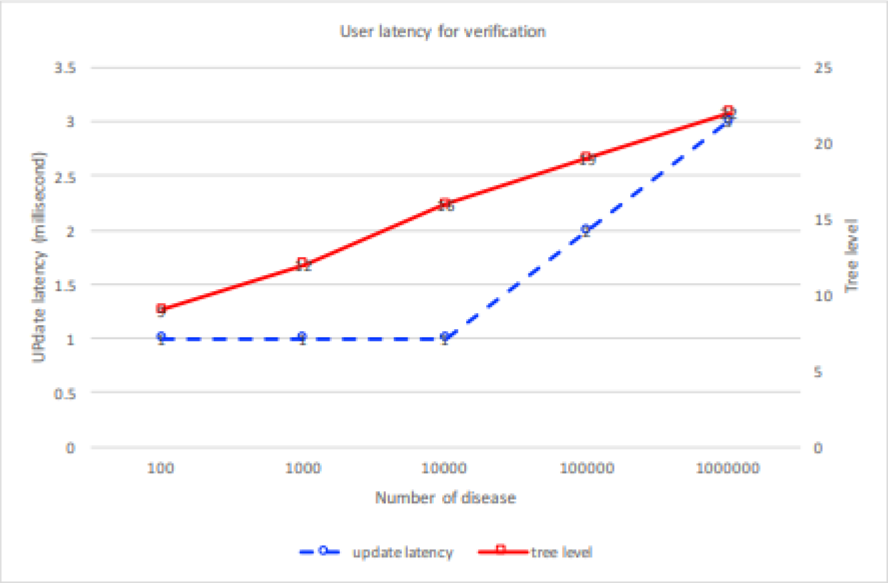


Figure 4.9: User verification latency and related *hdmt* tree level.

4.5Computational cost of the cloud computing

Downloading all of the leaf nodes and business lists using the SADS approach, the user downloads all of the leaf nodes from the cloud, as well as the related business list. The level of the tree and the quantity of the specified illness have a notable influence on the overall plant length. While the HDMT comprises 102, 103, 104, 105, 106, 107 disorders, we test the cloud lack of one illness. Fig. 4.10 illustrates the outcome. Even if there are 107 aberrant leaves in the HDMT, a computationally affordable approach to finding them on the cloud level is reasonable as long as a user has at least 1,000 diseases to be searched in the SADS methodology. The cloud looks for the higher-risk leaf node by utilising the key t1 = p f (k1, xa||...||xb). On the cloud, the cost of calculating a value is O (1).

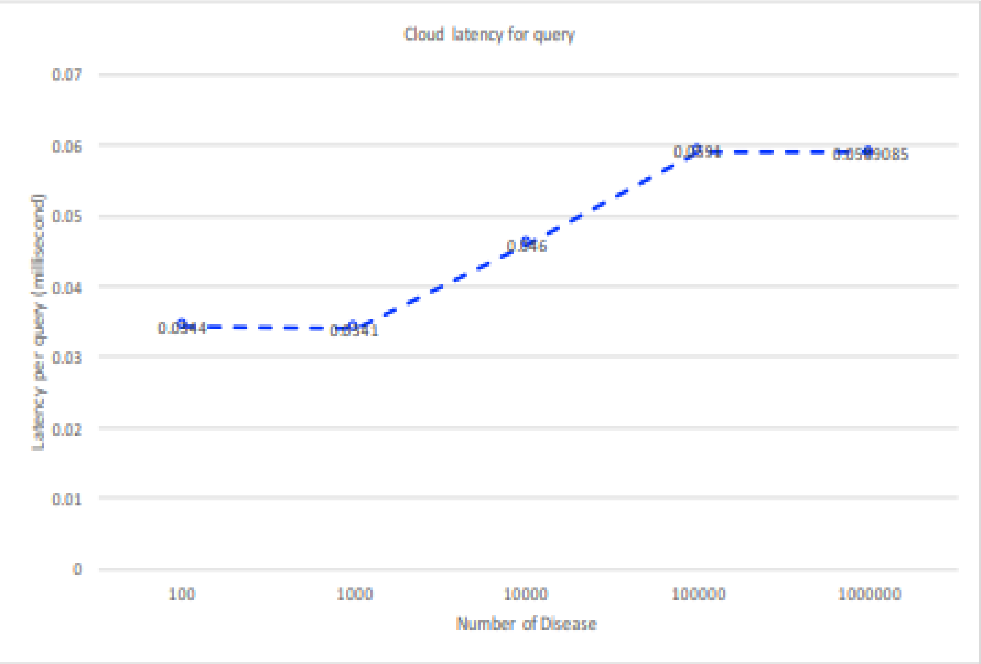


Figure 4.10: Cloud latency for the user to query one disease

4.5.1 Communication and its Cost

The hospital relies on integrated hardware and the encrypted illness threshold database to get the HDMT out to the patients. A single 32-bit node id is stored in each HDMT tree node. A single 256-bit key and a separate 256-bit value appears in each row of the encrypted data level. In the example in Table 4.3, when n diseases, n disease stages, and total L disease stages are given, the overall communication overhead for a hospital outsourcing the HDMT and encryption disease level is calculated as (2n n 2l0g n). The act is performed only once. Since, in this instance, it is safe for use at a hospital, it should be used in this location. To query diseases that can be discovered, the user uploads all of his or her sensory data when they register. When dealing with a user who has a certain number of senses, it is safe to assume that the result of the query is a list of ailments that are relevant to that person. Let's say each sensor and illness have a maximum of 32 bits; the sensors provide a data load of 32,000 bits, and the data retrieval for the query answer takes 32,000. When a user transfers an illness list to the cloud, it does a comprehensive scan of the whole HDMT, harvesting the information needed to find the leaf nodes and relevant partner lists. Downloading all the leaf nodes and the related list will take 32n (log D + 1) bits if the cloud HDDMT includes D illnesses and the user has only n to detect them. The fact is that the number "n" assigned to the user's ailment will be at least 1000. HDMT has determined that the disease number D will be roughly 107. Since this feature is available, the user is able to download only the leaf nodes and the associated administrative list for a fair cost.

Association rules mining is the application of data mining known as an association rule. Check out the TRANSACTION DATABASE of a successful trademark. The finding might be that the majority of bread buyers also purchase milk. Purchasing bread entails purchasing milk. In order to evaluate a candidate rule, two metrics are defined: guaranteed supply and expert guidance On the contrary, the number of trans transactions where both milk and bread are purchased divided by the number of transactions where bread is purchased is equal to the number of transactions where bread is purchased. The total number of transactions divided by the number of transactions when both brea and milk are purchased equals support. In order for a candidate to be considered an association rule, the person must have confidence and support.

Common item sets are found at the heart of traditional association rule mining approaches.

In order for both milk and bread to be often purchased in transactions, we are asserting that they are frequent transactions. You may easily find all association rules if you know all frequent itemsets.

To help us do this, we have carried out an investigation on the privacy measures needed while mining frequent item sets in a distributed manner. In other words, to hunt for frequent itemsets without telling either side, two or more parties are involved in a distributed database. A more detailed definition of the term "privacy" will be provided. By showing how our algorithms maintain privacy, we'll demonstrate how to deal with two fundamental approaches to data partitioning: vertical partitioning and horizontal partitioning. Based on our research, we believe that Lifton and his students were the first to examine privacy-preserving distributed mining of association rules/frequent itemsets. The following scientists developed an algebraic solution for vertically partitioned data: Vaidya and Lifton. In spite of this, the solution may lose many linearly arranged combinations of one pair's own exclusive sources of information to another. Furthermore, a decent one of frequent itemset's incremental profitability is quadratic in terms of number of transactions. When applying Yaw's generic secure compression algorithm as a subprotocol, Karantoglou and Lifton use Yaw's generically named algorithm as a protoco However, as stated by Gōldreið, trustworthy computation proto---bed strategies. Cost overruns can be more prevalent in data mining than in other industries due to the enormous magnitude of the input. The solution is only effective for three or more people, not two. For several difficulties, particularly those concerning mining association rules/frequent itemsets, several projects have made headway. Projects like these, in contrast, are focused on providing privacy for single transactions and concealing rules, rather than on securing the processing of distributed data.

The authors first tried to resolve this privacy-restricting decentralised mining issue by Lindell and Pinkas. However, their proposal only deals with the classifying problem (“describing classifying activities that are contained inside discrete subsets of categories”), and not the association rule problem.

Here, we learn that because of the concerns of privacy-preserving data mining, the problems of generically secure computing may be seen as an application of existing protective protocols. Nevertheless, the protocols utilised here are rather expensive, and as a result, our specific solutions must be at least 15% more efficient in order to meet our goals.

In this project, we describe the security and availability issues that revolve on mining frequent item sets in a way that is consistent with inventory-saving mining.

We offer algorithms that can process data that has been partitioned vertically or horizontally. We build algorithms with two degrees of flexibility for vertically partitioned data. Currently, the solutio Our algorithms are effective because their overall operational velocity is linked to the velocity of light in transit.

This strategy is more efficient for horizontally partitioned data.

Additionally, our technique may work for multiple groups, as long as three or more are present.

Following this is the rest of the project: we provide a problem formulation and property definitions. We also offer two-party algorithms, where the algorithm structure is divided vertically by data, and the privacy is categorised as weak or strong. It is called Statistical Disclosure Control (SDC) in the statistics field, and it is designed to help respondents maintain their anonymity. privacy-preserving data mining (PPDM).

PPDM is used to securely allow many data owners to have access to aggregated data sets while keeping their data private. Even though privacy-preserving techniques have “evolved in a remarkably autonomous manner,” several other research groups commonly use these approaches. Data flipping, suppression, noise addition, and k-anonymity are all themes in the SDC and PPDM literatures. PPDM means any privacy-preserving measures in this area (regardless of provenance). For the sake of preserving data, accounting approaches that save the data tend to decrease the representation accuity. Modifying data may have an impact on the quality of the outcomes.

In determining the value of translated material for analysis, one looks at its effectiveness. Reducing privacy threats while increasing data value is the biggest privacy problem. Privacy-preserving strategies may be roughly classified into two categories: those that help to lower the danger of disclosing personal identification or sensitive data, and those that enhance anonymity. When someone is identifiable by a single feature (such as a credit card number) or a combination of qualities, then they have their identity disassociated (e.g., age, gender, postcode, and job description). Finding out the value of a sensitive property is called a sensitive attribute discovery (e.g., a medical diagnostic). Statistics-origination data transformation methods (such as data swapping or noise addition) do not provide any form of payment

A calculation based on distance would take place, for example, in order to compute the quantity of proof. PPDM methods are referred to as privacy models due to the possibility of "retaining an undefined privacy or offering a privacy guarantee over the implemented data." Privacy may be ensured, for example, with k-anonymity. When each record in the data set is indistinguishable from at least k-1 other records, then the data set fulfils k-anonymity. We explore these techniques in detail. Usefulness may be quantified in several ways; for example, information loss may be estimated by calculating the differences between original and anonymized data, or usefulness may be geared toward a certain application.

The variety of data analysis methods and the requirement for specialised applications mean that assessments of data usefulness are more informative.

Personal privacy-preserving data mining techniques could be either very general or very specific.

The Generic method is a helpful tool for re-engineering data such that translated data may be utilised as input to do any data extraction task.

Without introducing new records, modifies databases so that fresh results can be provided

However, as a result, certain data mining algorithms provide outputs that are private, and techniques to shield this privacy (such as rule concealment) have been developed. Finally, for scenarios in which several data owners desire to derive insights from pooled data without risking their data positions, distributed shared-reserving systems are provided. At this piece, we take up the task of safeguarding health records stored in a medical facility; distributed data scenarios are outside the scope of our concern. Also, PPDM models and methods for detecting output privacy are specifically suited to particular data mining methodologies, which is why we do not specify them (and are not relevant to other data or processes mining algorithms).

CHAPTER 5

Methodology

5.1 Introduction

While several WBAN (Wireless Body Area Network) schemes and solutions have recently been introduced, they include the medium access protocole for WBAN, which is energy-efficient, and data forwarding infrastructure between biosensors. Due to the sensors' limited resources, the gathered data streams are typically not directly transferred to the healthcare centre. As illustrated in Fig. 5.1, sensors in each wearable health system deriodically collect users' physiologic data and communicate these data to their smartphones, where calculations are made and then sent back to the sensors. The smartphone unites previously processed medical data with previously processed medical data from the user, and then sends the combined data to a nearby WBN gateway.

A wide variety of people will utilise WBAN-gates to gather their personal items, such as medications. This hospital thereafter receives the medical supplies through the WBAN-gateway. For WBAN medical device health care systems, users must submit sensitive personal data, such as age, name, gender, and medical information, which raises major issues regarding leakage and data theft. The same holds true for the health centre, which refuses to share illness models with the public. There is the possibility that cell phones or WBANgateways will be infiltrated by burglars, who may eavesdrop on sensitive personal conversations and even collect confidential healthcare data. As a result, many privacy-protective measures are now available in the e-healthcare system. The validity and efficiency of privacy-preserving health-care systems that use encrypted data have to be carefully assessed. The remote e-healthcare system would have to address the following issues:

5.2 ChallengesTo Deal With

Challenges to privacy andsecurity. All of the users' physiologically and personally identifiable data, and the clinic's sickness models, should be protected. In order to protect the plaintext, it must be encrypted such that only the cryptotext can view it. It's also acceptable to assume that the attacker knows about users' personal information or even their medical conditions at times. The thief in this example will never be able to recover any other text included in the corresponding encrypted data. When using known-plaintext approaches, the system should be protected.

Challenge to accuracy. As a result, certain protective schemes that are based on entry-recorded data must establish user personal information and the disease-surveillance center's treatment of the condition to fulfil the safety criteria of the remote e-healthcare system. Accuracy of computation might be compromised if a standardisation effort is undertaken. Also, there are alternative methods, such as differential practise, which may add random value to commercial processes, which might cause medical disaster in some cases. as a result, the fickle rationale-restricting medical analysis remover e-healthare system must be accurate.

Challenges to efficiency.This component of the system model does not have interactive systems, and as a result, the time-consuming procedures featured in the previous chapter are usually required. Most economical approaches are founded on data entry utilising a large corporate overhead. More recent research confirms that privacy-preserving strategies such as multi-party privacy enhancing protocols work in an interactive setting, but cannot be employed in a non-interactive one. Due to this, the private remote e-healthcare system will have to find ways to improve efficiency. An efficient and privaпcy-preserving classfiпation on patient health data contained within a remove e-healthcare server, which enables authorised users the ability to partially transfer dat packets individually. There are WBN-attached metered packets in a non-interactive, nonservice-preserving method, utilising the packets' properties. This chapter includes several important contributions:

5.3 Methodology

- **First**,We give a demonstration of the PPC scheme, a no-interference non-interactive classification levelization scheme for customers' medical packets in WBAN-gateway applications. In particular, the WBAN-gateways rely on the legitimacy of the medical kits and trust the kits in a manner that serves as a resource.

- **Second,**To evaluate the system's performance, we use two Java server implementations and one Android application. The findings show that the proposed approach is effective, particularly in regard to computation complexity and communication requirements. In our proposed strategy, the confidentiality of the patients' personal information and the privacy of the sickness models will also be preserved, as well as the security of the medical facility, according to the study.

The foundation of the proposed method is to apply commutative encryption to all of the property's data components. Commutative encryption means that the order of encryption does not matter in many cryptographic techniques. Encryption keys E1 and E2, along with any message m, result in E1(E2(m)) being equal to E2(E1(m)). In both decryption and encryption, it is sufficient to use a single key, first decrypting a message encrypted with two keys before trying the process again. In essence, the core concept is that each resource encrypts its data set using its own private keys before sending the encrypted data set to the next source. This empfies the fact that it encrypts the received data and passes it on to the next node until the entire network has encrypted the data. The two sorts of data set attributes mentioned in the preceding section are called key attributes and non-key attributes. When two or more data sets from various sources are used in a computation, the methods used will be determined by the attributes of each data set. If and only if the original values are the same, the encrypted values of the key attribute will be the same in different data sets. This means that all different types of data sets, including ones that originate from many sources, may be united based on their primary properties, which are all encrypted. While concealing the true value of the key attribute, the encryption safeguards the integrity of both the source and the querying properties.

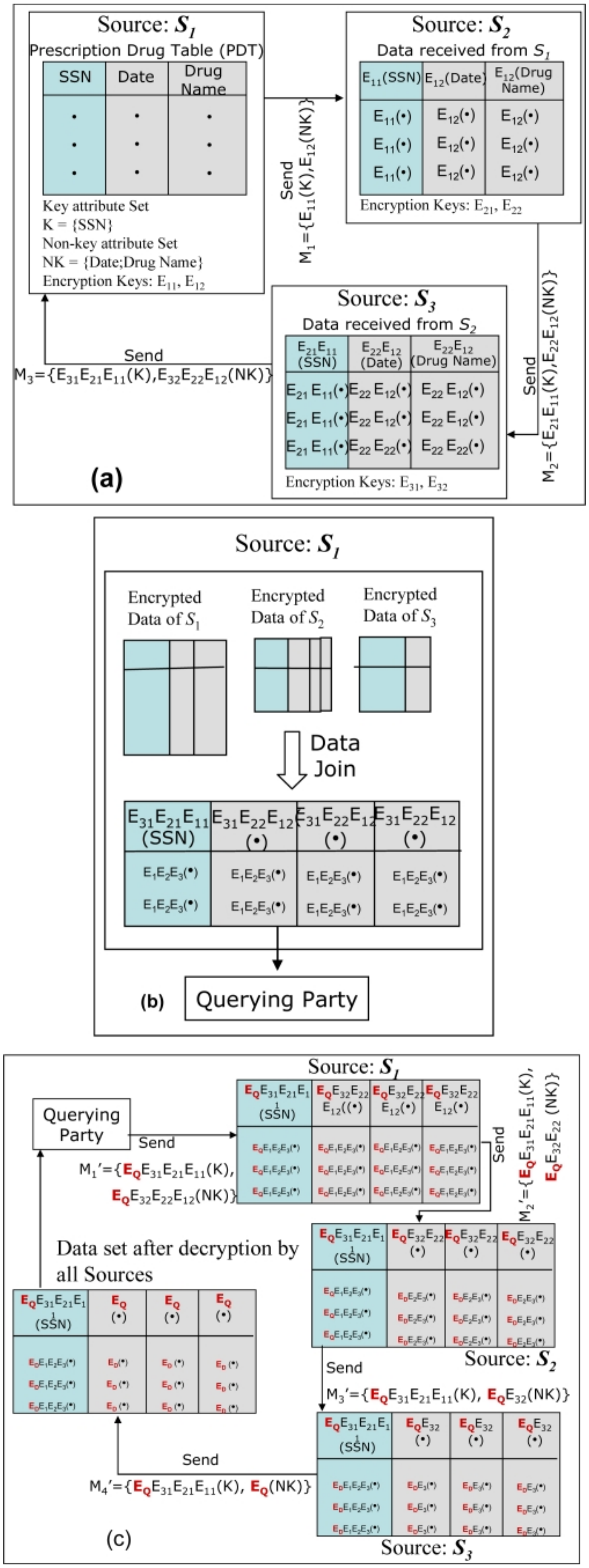
The algorithm consists of three phases: encryption, joining, and decryption. The next stages are as follows.

Encryption

When all data sets have been joined, all sources encrypt the joining attribute of each data set. Two encryption and decryption keys (Ei1, Di1) and (Ei1, Di2) are created for each Si that produces (Ei2, Di2). Ei1 and Ei2 respectively, are used to encrypt the values of the key attribute and non-key attributes of all data sets in the dataset (1in). Source Si encrypts the data, which it then delivers to its adjacent neighbour Sj, where j is (i+1) mod n. Next-door neighbour Sj uses its own encryption keys Ej1 and Ej2 to encrypt the incoming data set. It should be noted that Sj, in addition to setting the table order in the encrypted data set to prevent a holder of the data set from matching the encrypted values of the key attribute with their actual values, additionally permutes the tables in the encrypted data set to hinder this third party from establishing the value of the key attribute using their actual values. Next to Sj's neighbouring node, the resulting data set is transmitted to Sj's next neighbour for encryption. Until all of the sources have encrypted the data, the process will continue.

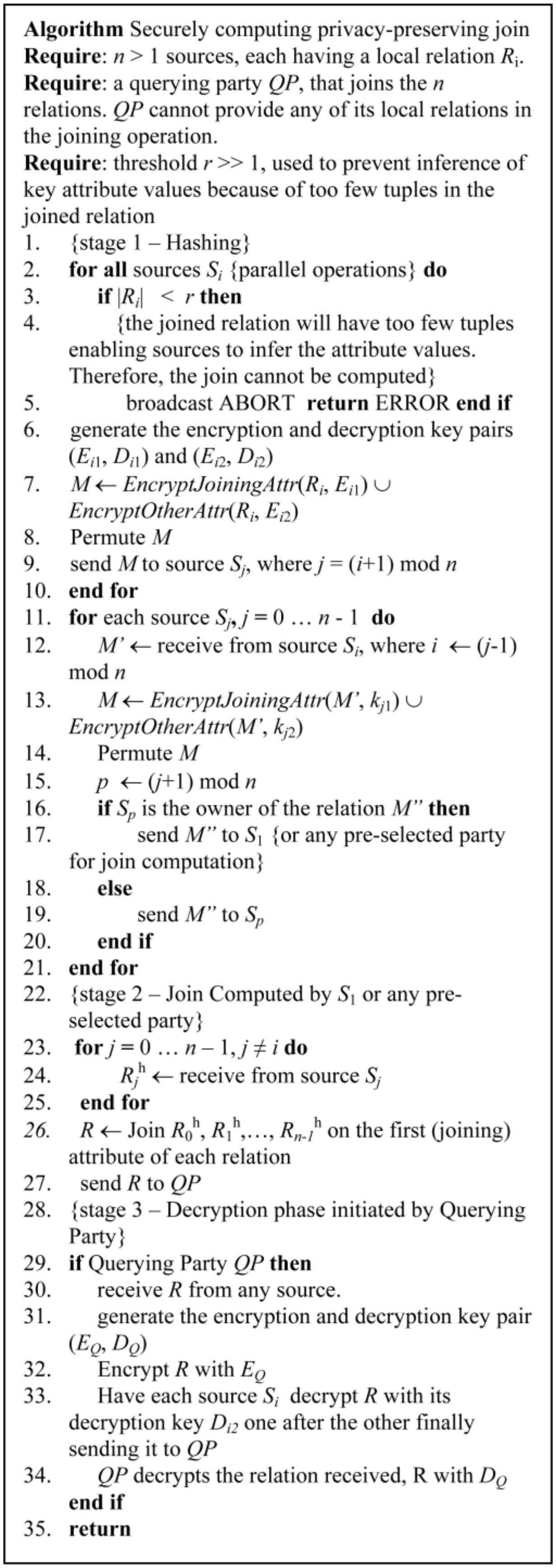
Decryption

In electronic format, the combined data set provided to the querying property has both key and non-key attributes. The querying property, on the other hand, should be able to acquire non-key attribute values without having to learn or infer the real value of the key attribute in any of the tuples. As a result, the merged data set's non-key attribute values must be decrypted sequentially using the decryption keys of all sources. The decoding of the non-key properties of the merged data set can only happen at each source since each has its unique encryption key that no one else knows. As previously said, each side must decode each other one at a time in order to receive the real data. On the other hand, the querying attribute must ensure that none of the sources may see the real data.


Source: S_1
Prescription Drug Table (PDT)
SSN
Date
Drug Name
Key attribute Set
K = {SSN}
Non-key attribute Set
NK = {Date;Drug Name}
Encryption Keys: E_{11}, E_{12}
Send
$M_1=\{E_{11}(K),E_{12}(NK)\}$
Source: S_2
Data received from S_1
E_{11}(SSN)
E_{12}(Date)
E_{12}(Drug Name)
$E_{11}(\bullet)$
$E_{12}(\bullet)$
Encryption Keys: E_{21}, E_{22}
Send
$M_2=\{E_{21}E_{11}(K),E_{22}E_{12}(NK)\}$
Source: S_3
Data received from S_2
$E_{21}E_{11}$ (SSN)
$E_{22}E_{12}$ (Date)
$E_{22}E_{12}$ (Drug Name)
E_{21} $E_{11}(\bullet)$
E_{22} $E_{12}(\bullet)$
E_{22} $E_{22}(\bullet)$
Encryption Keys: E_{31}, E_{32}
Send
$M_3=\{E_{31}E_{21}E_{11}(K),E_{32}E_{22}E_{12}(NK)\}$
(a)
Source: S_1
Encrypted Data of S_1
Encrypted Data of S_2
Encrypted Data of S_3
Data Join
$E_{31}E_{21}E_{11}$ (SSN)
$E_{31}E_{22}E_{12}$ $(\bullet)$
$E_{31}E_{22}E_{12}$ $(\bullet)$
$E_{31}E_{22}E_{12}$ $(\bullet)$
$E_1E_2E_3(\bullet)$
(b)
Querying Party
Source: S_1
Querying Party
Send
$M_1'=\{E_QE_{31}E_{21}E_{11}(K), E_QE_{32}E_{22}E_{12}(NK)\}$
$E_QE_{31}E_{21}E_1$ (SSN)
$E_QE_{32}E_{22}$ $E_{12}((\bullet)$
$E_QE_{32}E_{22}$ $E_{12}(\bullet)$
$E_QE_1E_2E_3(\bullet)$
Send
$M_2'=\{E_QE_{31}E_{21}E_{11}(K), E_QE_{32}E_{22}(NK)\}$
$E_QE_{32}E_{22}$ $(\bullet)$
$E_QE_2E_3(\bullet)$
Source: S_2
Send
$M_3'=\{E_QE_{31}E_{21}E_{11}(K), E_QE_{32}(NK)\}$
E_QE_{32} $(\bullet)$
$E_QE_3(\bullet)$
Source: S_3
Send
$M_4'=\{E_QE_{31}E_{21}E_{11}(K), E_Q(NK)\}$
Data set after decryption by all Sources
E_Q $(\bullet)$
$E_Q(\bullet)$
(c)

The technique displayed in the image behind this paragraph may be used to obtain the join of various disparate data sets across several sources without disclosing any private or confidential information to the inquiring party. The party requesting information cannot be the source of any merged data set that will be used to analyse the query. As the querying party, if it is one of the data sources, it can use the non-key properties that the joined data set and its local data set share to calculate the real value of the identification (key) property. The sample query in the introductory section is an excellent illustration of this concept.

Using criteria like Drug Name and Prescription Date, the chemist can identify the identities of HIV-positive individuals. By extension, as inquiries that require the merging of heterogeneously dispersed data with the asking party as one of the data sources are incompatible with this technique, so are inquiries that include heterogeneous dispersions of data with a query subject as a data source. Also, there is a concern about how to handle link data sets that are too tiny. The join is not calculated if either of the input data sets has less records than a given threshold value r.



Algorithm Securely computing privacy-preserving join
Require: $n > 1$ sources, each having a local relation R_{i}.
Require: a querying party *QP*, that joins the n relations. *QP* cannot provide any of its local relations in the joining operation.
Require: threshold $r >> 1$, used to prevent inference of key attribute values because of too few tuples in the joined relation

```
{stage 1 – Hashing}
for all sources $S_i$ {parallel operations} do
    if $|R_i| < r$ then
        {the joined relation will have too few tuples enabling sources to infer the attribute values. Therefore, the join cannot be computed}
        broadcast ABORT  return ERROR end if
    generate the encryption and decryption key pairs ($E_{i1}$, $D_{i1}$) and ($E_{i2}$, $D_{i2}$)
    $M \leftarrow EncryptJoiningAttr(R_i, E_{i1}) \cup EncryptOtherAttr(R_i, E_{i2})$
    Permute $M$
    send $M$ to source $S_j$, where $j = (i+1) \bmod n$
end for
for each source $S_j$, $j = 0 \ldots n - 1$  do
    $M' \leftarrow$ receive from source $S_i$, where $i \leftarrow (j-1) \bmod n$
    $M \leftarrow EncryptJoiningAttr(M', k_{j1}) \cup EncryptOtherAttr(M', k_{j2})$
    Permute $M$
    $p \leftarrow (j+1) \bmod n$
    if $S_p$ is the owner of the relation $M''$ then
        send $M''$ to $S_1$ {or any pre-selected party for join computation}
    else
        send $M''$ to $S_p$
    end if
end for
{stage 2 – Join Computed by $S_1$ or any pre-selected party}
 for $j = 0 \ldots n - 1, j \neq i$ do
        $R_j^{\mathrm{h}} \leftarrow$ receive from source $S_j$
  end for
  $R \leftarrow$ Join $R_0^{\mathrm{h}}, R_1^{\mathrm{h}}, \ldots, R_{n-1}^{\mathrm{h}}$ on the first (joining) attribute of each relation
  send $R$ to $QP$
{stage 3 – Decryption phase initiated by Querying Party}
if Querying Party $QP$ then
    receive $R$ from any source.
    generate the encryption and decryption key pair ($E_Q$, $D_Q$)
    Encrypt $R$ with $E_Q$
    Have each source $S_i$ decrypt $R$ with its decryption key $D_{i2}$ one after the other finally sending it to $QP$
    $QP$ decrypts the relation received, R with $D_Q$
end if
return
```

5.4Prediagnosiswith Privacy Preservastion

In order to aid in prediagnosis, a health benefit user provides their own medical information to the prediction service, which includes a feature vector. To better protect the privacy of healthcare consumers, many initiatives have recently been launched. In 2014, Bos et al. constructed a cloud server utilising individualised encryption for private information services. Lattisep-based homomorphophisches Kryrosysteme kann wirksam sichern, welche Daten medizinischen Patienten im Kontext zugeordnet sind. the prediagnostic model is overlooked in Bos et als, which is to say that everyone is aware of the model (including patients). Due to this, Bos et al's analysis had to be performed in a non-secure environment. High-level security was reached, which means both the feature space's privacy and the predictive model's secret were protected.

In this example, the service provider supplies the prediagnostic model, while the patient supplies the feature vector. Information of both parties should remain confidential. Three major privacy-preserving classifiers (hypermoodle, neural networks, and decision trees) were created using the previous cryptosystem. A unique conditional oblivious transfer operator was utilised to develop an efficient privacy-preserving classifier, as reported by Wu et al. Multivariate polynomials and multivariant transprotoalogy were used by Jia et al. to develop an SVM that guarantees anonymity. To this end, the researchers designed and implemented a Jia et al. procedure. The procedure eliminated the need for time-consuming homomorphic encryption, yet provided excellent communication and financial transaction efficiency. Based on the nonlinear kernel SVM, the article published by Zhu et al. utilised lightweight multiparty random marking and polynomial aggregation techniques to produce a medical prediagnosis system using the uncoordinated kernelRVM. In the paper Zhu et al., privacy of the user's feature vector is guaranteed, but the secrecy of the SVM classifier is kept secret. Several K-means clustering methods that preserve privacy have been described.

Under differential private use, clustering has also been investigated. The two biggest challenges faced by independent-purpose K-means are budget allocation and initial centroids selection. The two distinctive ways to determining the number of iterations include having fixed iterations. Updated K-means clustering techniques are guaranteed to provide differential confidentiality. To assess the MSE between noisy and real centroids, the researchers came up with three ways for estimating the number of iterations and budget allocation. Some techniques, such as random selection, division of data into equal sized sections, and locating the central points, exist for the first selection. It pro-posed a DPLK-based algorithm which incorporates several policies, and which results in better selection of beginning point locations through process improvement. The Kmeans method is used to each subgroup separated by the original dataset. Computational efficiency may be increased by cluster analysis projects that employ distributed computing platforms. MapReduce, a fundamental yet effective parallel programming paradigm, was used to construct a parallel K-means clustering technique. To assign each sample to the nearest centre, the M-function is employed. To update the new centres, the R-function is utilised.

The Internet of Medical Things (IoMT) is a network of terminals and linked devices that deliver high-quality medical services. The analysis of health data for disease diagnosis and prediction is an important application in IoMT. The data owner (DO), the cloud service provider (CSP), the data processing centre (DPC), and the data consumer are the four main players in the Internet of Medical Things (DU).

(1) D: Wear terminal devices with health sensors that will send their health data to the cloudservice provider. DOs are presumed to be trustworthy in the system.

(2) S: PLoad all data in the cloud stores. S is believed to be telling the truth.

(3) D: The data processing facility processes the information and provides the findings to the users. In the system, DPC is regarded as reliable. D.U. : The data owner gets data computing results from the data distribution centre and uses the information to conduct duties such as diagnosis and prediction. However, the analysis process and release of the analysis findings may result in the leakage of users' privacy information. In order to preserve privacy in the k-means clustering algorithm, we construct a clustering algorithm that fulfils the differential privacy. In a distributed environment — Map Reduce framework — we deploy our algorithm forcomputational efficiency. In a distributed system, a clustering method that fulfils the differential privacy is also conceivable..

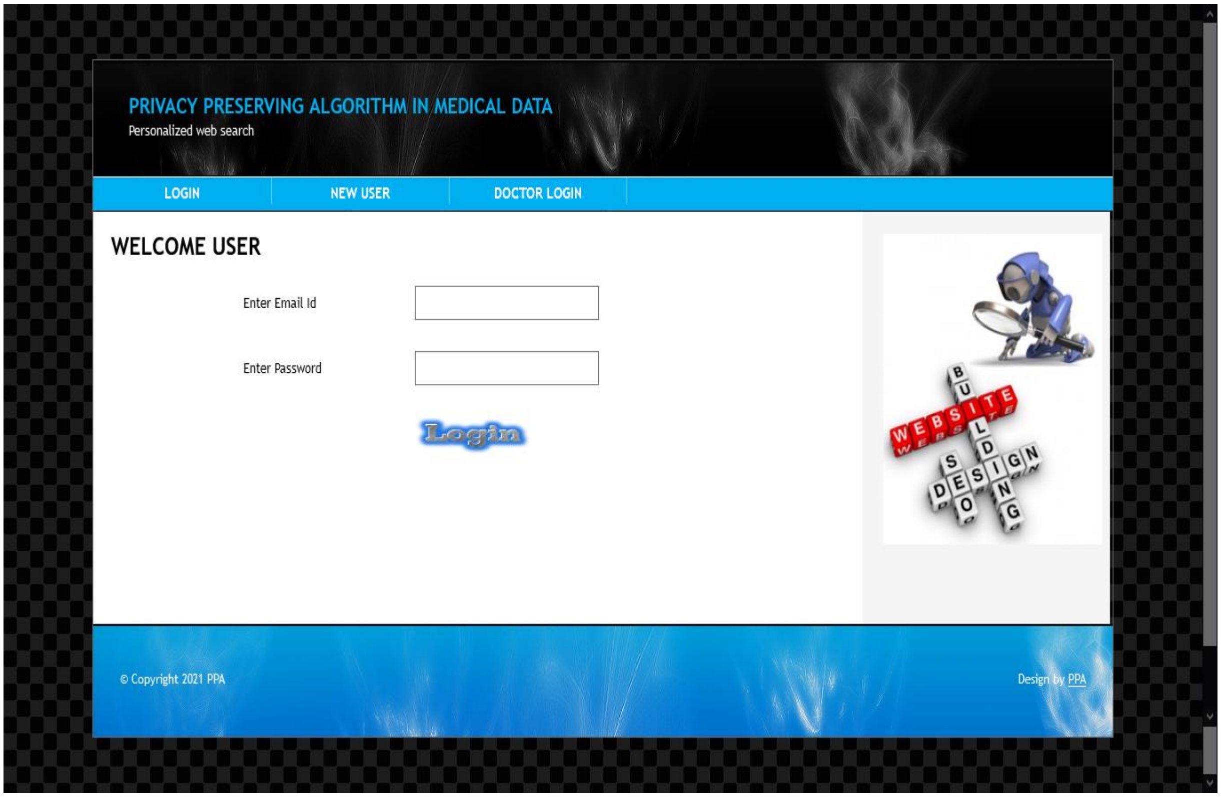
CHAPTER 6

1. Implementation and Experiment

This project divided in following module:

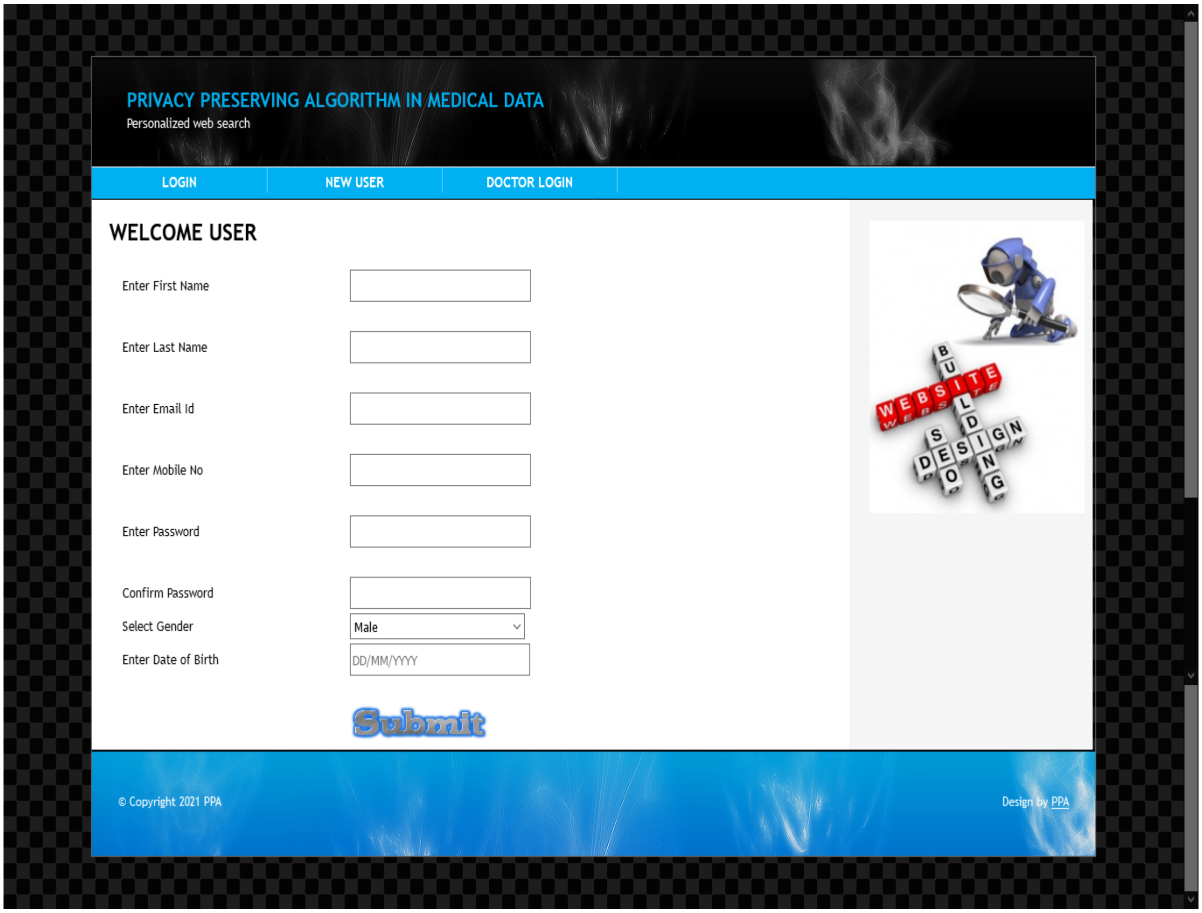
1. Admin Login:

In this module user can enter email id and password for authentication and authorization to access web sources of medical dataset to gain access to the stored information.



2: User Registration:

An enrolled client is a customer who has just signed up for a site, programme, or other framework. Enlisted clients demonstrate their character by signing in and supplying credentials (such as a username or email address, as well as a secret key) to the framework. Clients can enlist by selecting a register or sign up job and giving these certificates to the initial go through in frameworks designed for general use. Clients who have enlisted may be eligible for perks not available to unregistered customers.



PRIVACY PRESERVING ALGORITHM IN MEDICAL DATA

Personalized web search

WELCOME USER

Enter First Name	
Enter Last Name	
Enter Email Id	
Enter Mobile No	
Enter Password	
Confirm Password	
Select Gender	Male
Enter Date of Birth	DD/MM/YYYY

© Copyright 2021 PPA

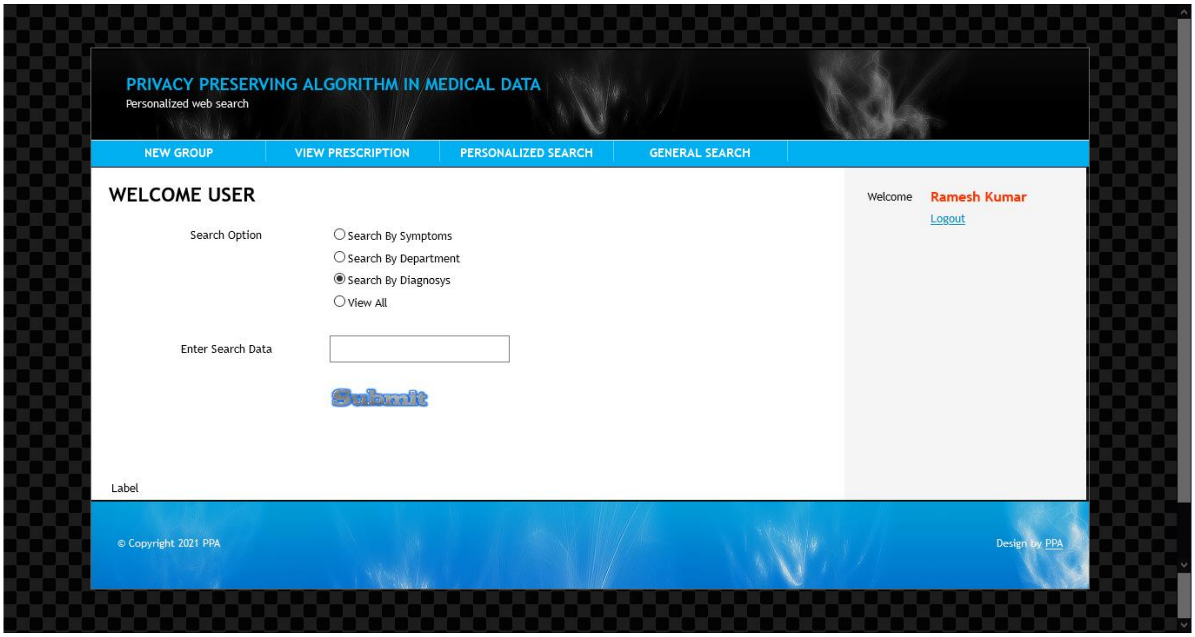
Design by PPA

3. **General Search Module**

Users can search and obtain public medical data of other users in this module. Other users have the option of making medical data public or private, depending on their preferences.

4. **View Prescription:**

Users can view information on their own therapy and store it using public, private, and group search options. Users can create new groups to add more people to that group.



PRIVACY PRESERVING ALGORITHM IN MEDICAL DATA
Personalized web search

WELCOME USER

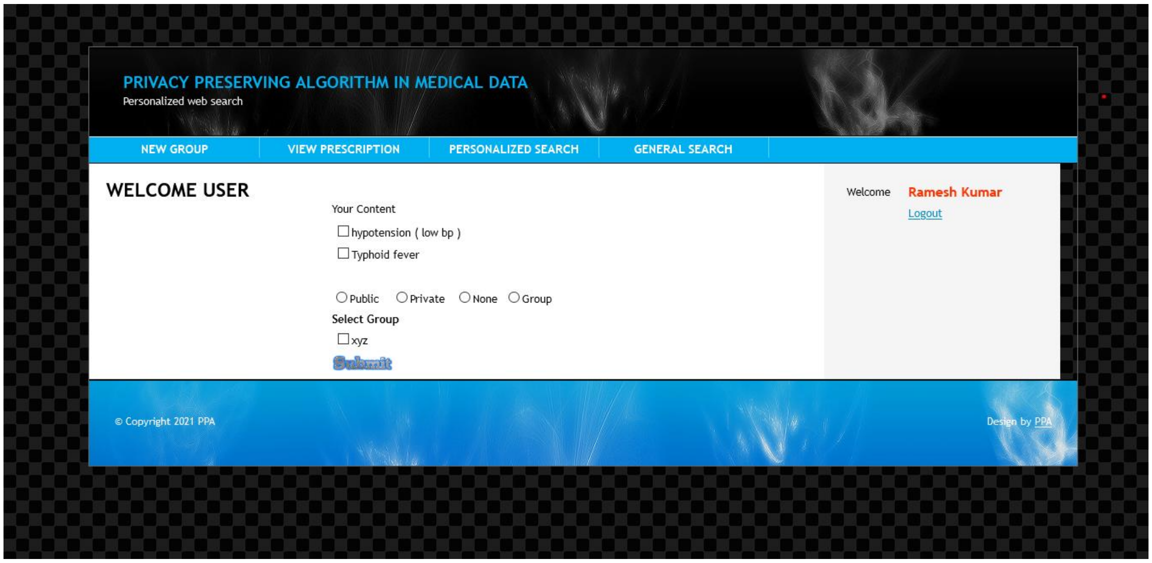
Welcome Ramesh Kumar

Search Option
○ Search By Symptoms
○ Search By Department
● Search By Diagnosys
○ View All

Enter Search Data

Label

© Copyright 2021 PPA
Design by PPA



PRIVACY PRESERVING ALGORITHM IN MEDICAL DATA
Personalized web search

WELCOME USER

Welcome Ramesh Kumar

Your Content
□ hypotension (low bp)
□ Typhoid fever

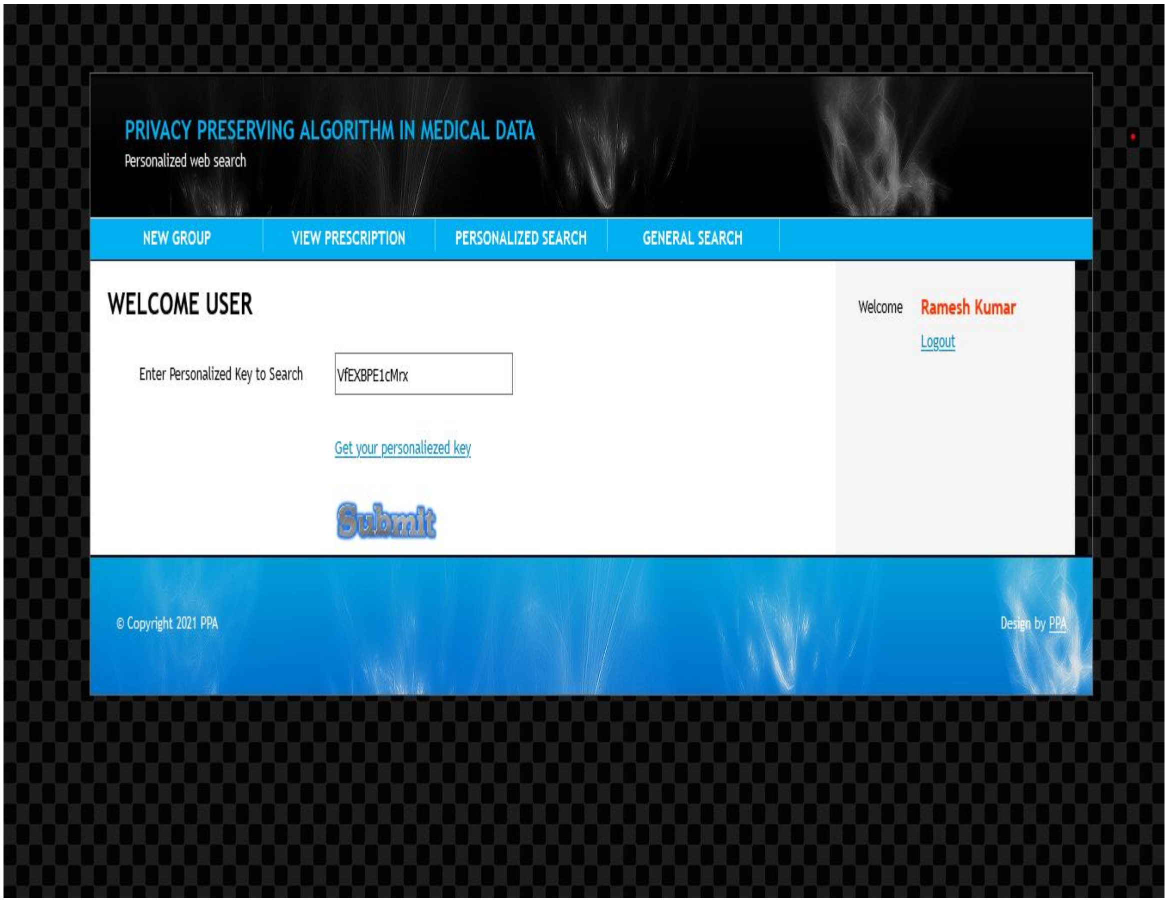
○ Public ○ Private ○ None ○ Group

Select Group
□ xyz

© Copyright 2021 PPA
Design by PPA

5. Private Search:

Aside from producing their own private medical data, the user may search the private medical data of other users. That user must have a valid secret in order to produce a private key for themselves. Without private, the user is unable to look for any confidential medical information.



PRIVACY PRESERVING ALGORITHM IN MEDICAL DATA

Personalized web search

WELCOME USER

Enter Personalized Key to Search: VfEXBPE1cMrx

Get your personaliezed key

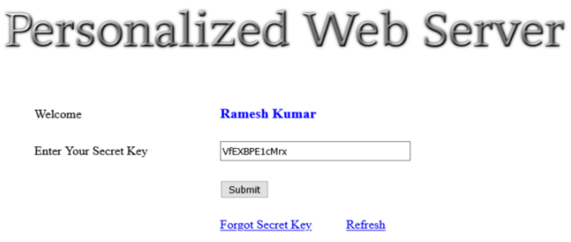
Welcome Ramesh Kumar

© Copyright 2021 PPA

Design by PPA

6. Personalized Web Server

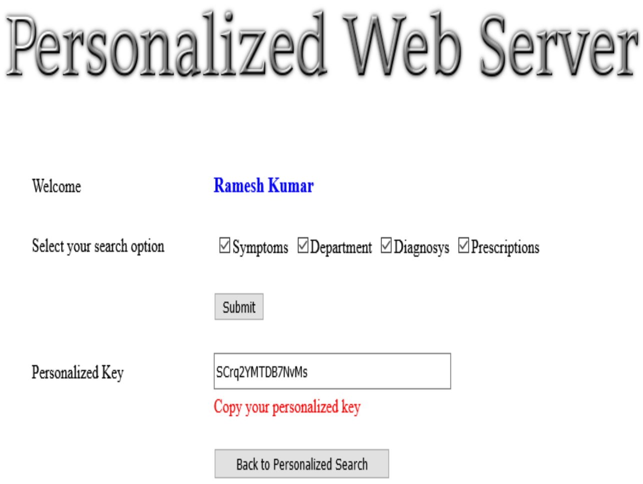
It's what's utilised to produce a private key. This is a third-party server where the user can produce private keys for himself with the use of a legitimate secret key.



Personalized Web Server

Welcome Ramesh Kumar

Enter Your Secret Key: VfEXBPE1cMrx



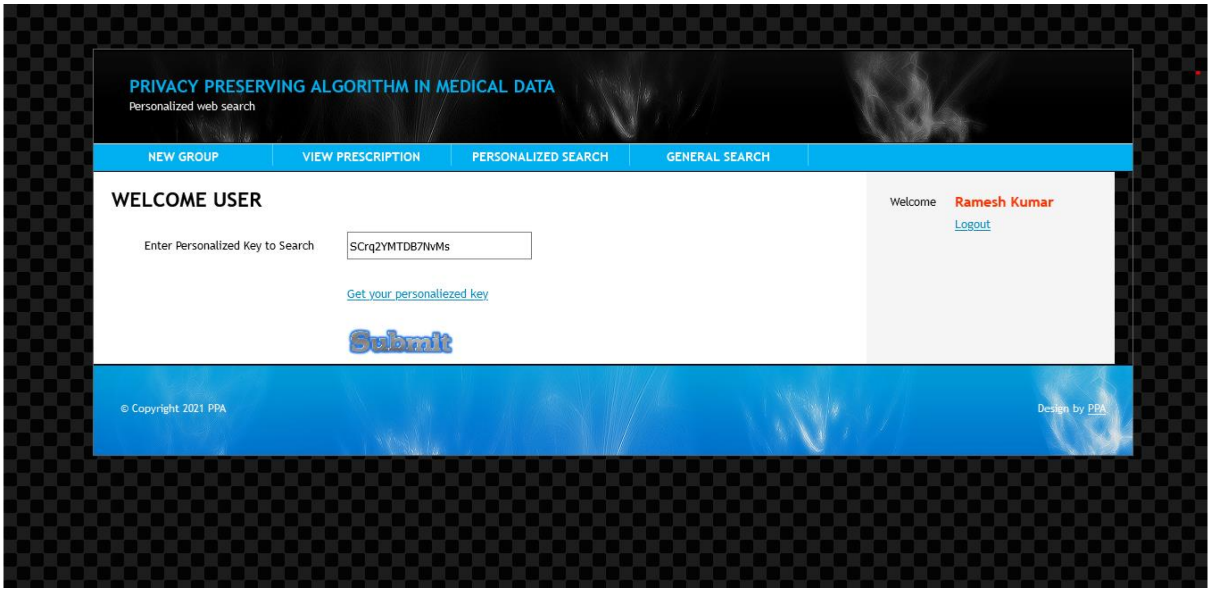
Personalized Web Server

Welcome Ramesh Kumar

Select your search option ☑Symptoms ☑Department ☑Diagnosys ☑Prescriptions

Personalized Key: SCrq2YMTDB7NvMs

Copy your personalized key



PRIVACY PRESERVING ALGORITHM IN MEDICAL DATA

Personalized web search

WELCOME USER

Enter Personalized Key to Search: SCrq2YMTDB7NvMs

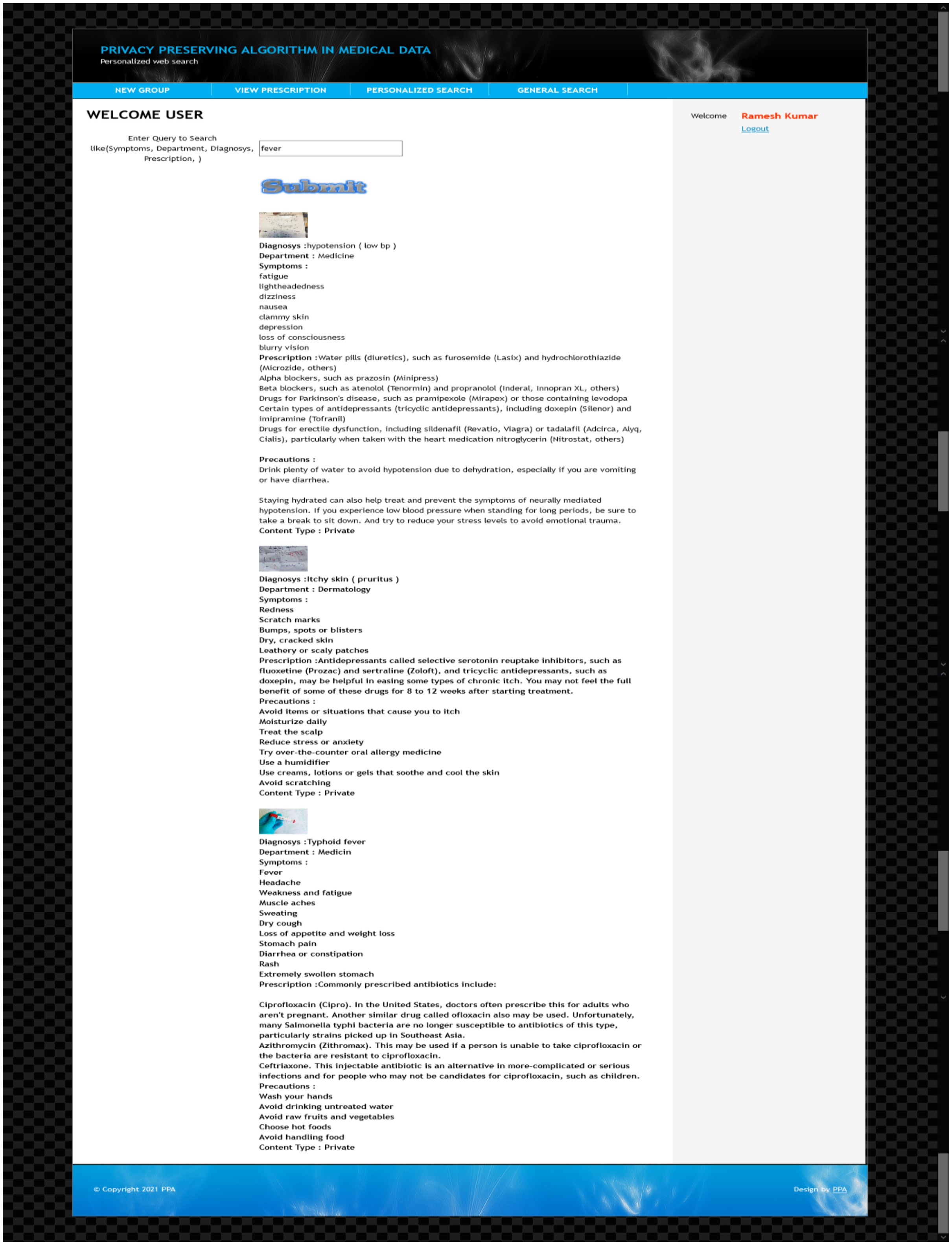
Welcome Ramesh Kumar

© Copyright 2021 PPA

Design by PPA

7. Private Data Search Module

The user can search the medical dataset's encrypted private data here. He'll need a private key for this, which he can get from the server.



WELCOME USER

Enter Query to Search like(Symptoms, Department, Diagnosys, Prescription,) fever

Diagnosys :hypotension (low bp)
Department : Medicine
Symptoms :
fatigue
lightheadedness
dizziness
nausea
clammy skin
depression
loss of consciousness
blurry vision
Prescription : Water pills (diuretics), such as furosemide (Lasix) and hydrochlorothiazide (Microzide, others)
Alpha blockers, such as prazosin (Minipress)
Beta blockers, such as atenolol (Tenormin) and propranolol (Inderal, Innopran XL, others)
Drugs for Parkinson's disease, such as pramipexole (Mirapex) or those containing levodopa
Certain types of antidepressants (tricyclic antidepressants), including doxepin (Silenor) and imipramine (Tofranil)
Drugs for erectile dysfunction, including sildenafil (Revatio, Viagra) or tadalafil (Adcirca, Alyq, Cialis), particularly when taken with the heart medication nitroglycerin (Nitrostat, others)

Precautions :
Drink plenty of water to avoid hypotension due to dehydration, especially if you are vomiting or have diarrhea.

Staying hydrated can also help treat and prevent the symptoms of neurally mediated hypotension. If you experience low blood pressure when standing for long periods, be sure to take a break to sit down. And try to reduce your stress levels to avoid emotional trauma.
Content Type : Private

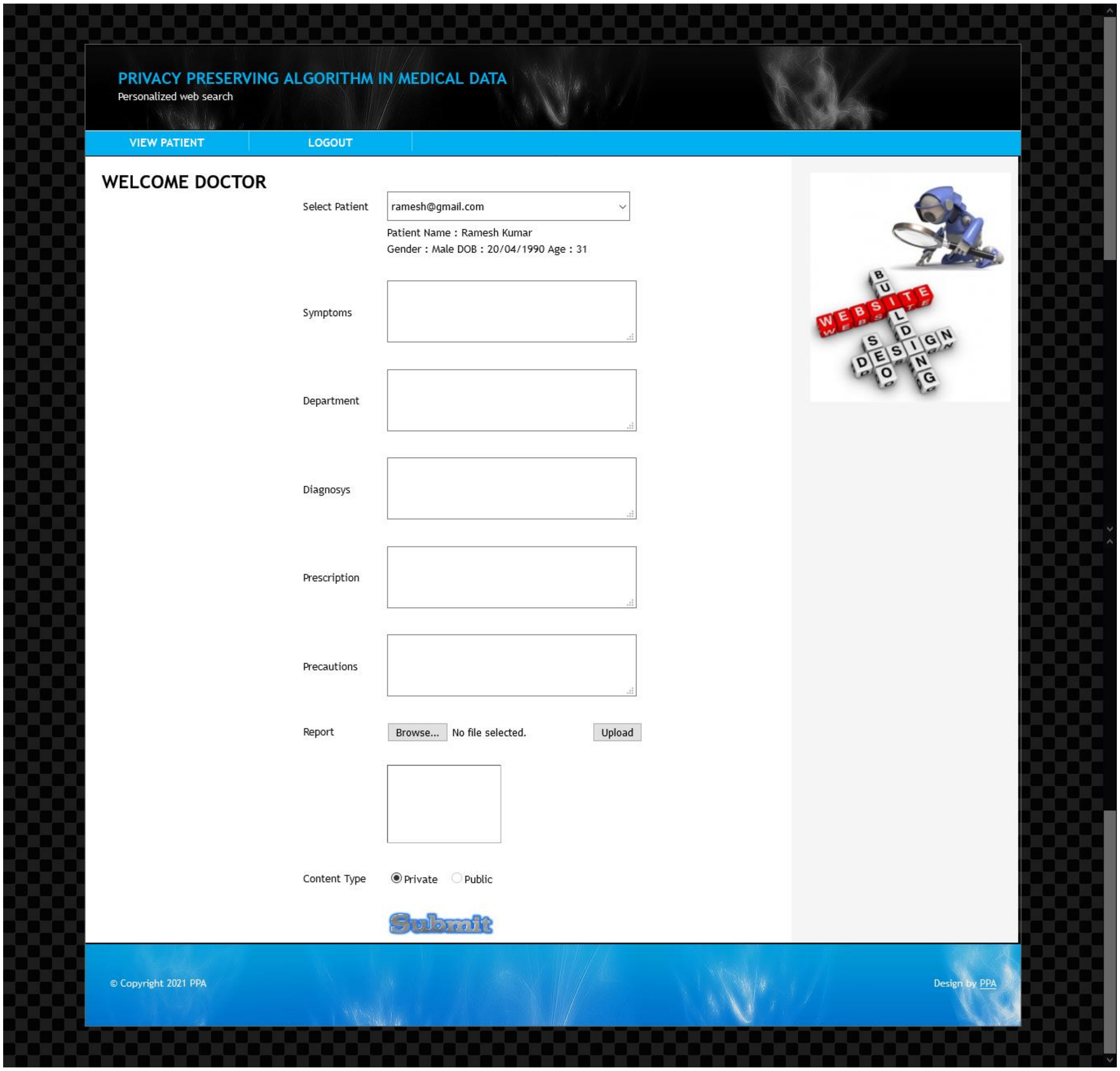
Diagnosys :Itchy skin (pruritus)
Department : Dermatology
Symptoms :
Redness
Scratch marks
Bumps, spots or blisters
Dry, cracked skin
Leathery or scaly patches
Prescription :Antidepressants called selective serotonin reuptake inhibitors, such as fluoxetine (Prozac) and sertraline (Zoloft), and tricyclic antidepressants, such as doxepin, may be helpful in easing some types of chronic itch. You may not feel the full benefit of some of these drugs for 8 to 12 weeks after starting treatment.
Precautions :
Avoid items or situations that cause you to itch
Moisturize daily
Treat the scalp
Reduce stress or anxiety
Try over-the-counter oral allergy medicine
Use a humidifier
Use creams, lotions or gels that soothe and cool the skin
Avoid scratching
Content Type : Private

Diagnosys :Typhoid fever
Department : Medicin
Symptoms :
Fever
Headache
Weakness and fatigue
Muscle aches
Sweating
Dry cough
Loss of appetite and weight loss
Stomach pain
Diarrhea or constipation
Rash
Extremely swollen stomach
Prescription :Commonly prescribed antibiotics include:

Ciprofloxacin (Cipro). In the United States, doctors often prescribe this for adults who aren't pregnant. Another similar drug called ofloxacin also may be used. Unfortunately, many Salmonella typhi bacteria are no longer susceptible to antibiotics of this type, particularly strains picked up in Southeast Asia.
Azithromycin (Zithromax). This may be used if a person is unable to take ciprofloxacin or the bacteria are resistant to ciprofloxacin.
Ceftriaxone. This injectable antibiotic is an alternative in more-complicated or serious infections and for people who may not be candidates for ciprofloxacin, such as children.
Precautions :
Wash your hands
Avoid drinking untreated water
Avoid raw fruits and vegetables
Choose hot foods
Avoid handling food
Content Type : Private

8. Doctor’s Prescription Module

Doctors save information regarding the patient's condition and treatment in this area. The user may access his account and view his medical information.



PRIVACY PRESERVING ALGORITHM IN MEDICAL DATA
Personalized web search

WELCOME DOCTOR

Select Patient: ramesh@gmail.com
Patient Name : Ramesh Kumar
Gender : Male DOB : 20/04/1990 Age : 31

Symptoms

Department

Diagnosys

Prescription

Precautions

Report: No file selected.

Content Type: ● Private ○ Public

© Copyright 2021 PPA

Design by PPA

9. Database Section

A data set is a logically organised collection of information that is often kept and retrieved electronically via a computer. Formal planning and demonstration approaches are typically employed to evolve knowledge bases when they are more unpredictable. The data base administration framework (DBMS) is a solution that captures and analyses data by collaborating with end users, apps, and the data set itself. The central offices that administer the data set are also included in the DBMS code. The whole information base, database management system, and related applications are referred to as a "data set framework." The word "data set" is commonly used to refer to a database management system (DBMS), an information base framework, or an application that is linked to a data set. PC researchers can organise information base administration frameworks based on the data set models that they support. Social data sets were increasingly common in the 1980s. The great majority of these models use SQL to construct and query data, and they represent data as lines and segments in a sequence of tables. Non-social data sets were popular in the 2000s, and they're referred to as No SQL since they don't utilise SQL.

CHAPTER 7

7.1 Conclusions

We proposed a new technique for merging and extracting data in a secure and privacy-preserving manner in this work. This suggested system employs a kind of encryption that cryptographically encodes each data set, including all of the qualification data from each source's attribute values, and then these sets are made mutually exclusive by utilising each source's own cryptographic keys. In order for C commutative encryption to ensure that encrypted keys from multiple data sets are the same, the keys' "origins" (source) should be as near to possible. The feature ensures that no source or querying property is allowed to collect information about an individual's private life, or any private information in the combined data set. There is no access to the final result set by anybody but the enquiring party owing to commutative encryption. Integrating data into a privacy-preserving individual without a complete identification will be a key research challenge in the future.

To overcome the trust barrier that prohibits patients from providing their medical data to the NHIN for research reasons, this strategy might be beneficial. Central collectors do not compel data owners to give their raw data to them. As a result, there is a significant risk of data aggregator security breaches because the system does not rely on a trusted middleman for any joining. In order for this method to execute successfully, a network of standards-based, high-throughput communication channels must be established among all participating parties in a federated system.

Firstly,In our pre-clinical research, we have devised a strategic protocol that is both successful and discreet. In the context of the discourse, there are two primary approaches. It uses an ideal PPCP to minimise risks for users, allowing them to access predication services from a service provider with little risks. It is moreover an efficient private-reserving healthcare suggestion service since it uses a unique encryption mechanism.

Secondly, we have given an efficient and reserve-saving health query across outdated, outdated network infrastructure, which keeps the user's personal resources available and ensures the security of the health service provider's assessment GooC has suddenly found the proposed sensor invention, which provids the discoverage of the high risk disease. Finally, the user queries the survival resulsat diagnosis logic.

Thirdly, we've designed an efficient privacy-preserving priority classification (PPCscheme) on patient healthcare data inside a remote e-Healthcare system. The suggested PPCscheme achieves priority classification and packets relay tasks, while preserving the privacy of the clients and the confidentiality of the healthcare's disease models.

7.2 Future work

We proposed a new technique for merging and extracting data in a secure and privacy-preserving manner in this work. This suggested system employs a kind of encryption that cryptographically encodes each data set, including all of the qualification data from each source's attribute values, and then these sets are made mutually exclusive by utilising each source's own cryptographic keys. In order for C commutative encryption to ensure that encrypted keys from multiple data sets are the same, the keys' "origins" (source) should be as near to possible. This feature ensures that any source or querying property, along with all of the source and querying properties' associated data, cannot access personally identifiable or private information from the aggregated data set. There is no access to the final result set by anybody but the enquiring party owing to commutative encryption. Integrating data into a privacy-preserving individual without a complete identification will be a key research challenge in the future.

To overcome the trust barrier that prohibits patients from providing their medical data to the NHIN for research reasons, this strategy might be beneficial. Central collectors do not compel data owners to give their raw data to them. As a result, there is a significant risk of data aggregator security breaches because the system does not rely on a trusted middleman for any joining. In order for this method to execute successfully, a network of standards-based, high-throughput communication channels must be established among all participating parties in a federated system.

List of References

Journal papers:

[1]Achieving data utility-privacy tradeoff in Internet of Medical Things: A machine learning approach Zhitao Guan a , Zefang Lv, Xiaojiang Du, Longfei Wu , Mohsen Guizani.

[2] Generation and evaluation of privacy preserving synthetic health data Andrew Yalea,∗ , Saloni Dashc , Ritik Dutta d, Isabelle Guyonb , Adrien Pavao b , Kristin P. Bennett.

[3] Gao,C., Zhang, X. & Liu, H. Data and knowledge-driven named entity recognition for cyber security.*Cybersecur* **4,** 9 (2021). https://doi.org/10.1186/s42400-021-00072-y.

Conference papers:

[1] B. Luca and X. Li, ''On differentially private longest increasing subsequence computation in data stream,'' Trans. Data Privacy, vol. 9, no. 1.

[2] C. F. D. Control, Prevention et al., "Hipaa privacy rule and public health. Guidance from cdc and the us department of health and human services," MMWR: Morbidity and mortality weekly report.

[3] J.Huang, M. Sharaf, and C. T. Huang, ''A hierarchical framework for secure and scalable EHR sharing and access control in multi-cloud,'' in Proc. ICPPW IEEE, Sep. 2012.

[4] M.J. Atallah,M. Blanton, and K. B. Frikken, ''Dynamic and efficient key management for access hierarchies,'' ACM Trans. Inf. Syst. Secur., vol. 12.

[5] M. C. Mont, P. Bramhall, and K. Harrison, ''A flexible role-based secure messaging service: Exploiting IBE technology for privacy in health care,'' in Proc. 14th Int. Workshop Database Expert Syst. Appl., Sep. 2003.

[6] P. Huang, B. Li, L. Guo, Z. Jin, and Y. Chen, “A robust and reusable ecgbased authentication and data encryption scheme for ehealth systems,” in Global Communications Conference (GLOBECOM), 2016 IEEE. IEEE,2016.

[7]S. Eberz, N. Paoletti, M. Roeschlin, M. Kwiatkowska, I. Martinovic, and A. Patané, “Broken hearted: How to attack ECG biometrics,” in NDSS Symposium 2017. Internet Society,2017.

[8]S. Sabitha and M. S. Rajasree, ‘‘Anonymous-CPABE: Privacy preserved content disclosure for data sharing in cloud,’’ in Architecture of Computing Systems—ARCS. Cham, Switzerland: Springer, 2015.

[9] Y. Yin, Y. Zeng, X. Chen, Y. Fan, The internet of things in healthcare: an overview, J. Ind. Inf. Integr. 1 (2016) 3–13.

[10]M.Orsini, M. Pacchioni, A. Malagoli, G. Guaraldi, My smart age: An innovative mobile and ioMT framework for patient’s empowerment, in: Proc. IEEE International Forum on Research and Technologies for Society and Industry.

APPENDIX

Important code logics

Libraries used:

```
 System;
 System.Collections.Generic;
 System.Linq;
 System.Web;
 System.Web.UI;
 System.Web.UI.WebControls;
```

Session establish

```
publicpartialclassLogin : System.Web.UI.Page
{
protectedvoid Page_Load(object sender, EventArgs e)
    {

    }
Users user = newUsers();
protectedvoid ImageButton1_Click(object sender, ImageClickEventArgs e)
    {
        user.EmailId =TextBox1.Text ;
        user.Password =TextBox2.Text ;
if (user.CheckEmailId())
        {
if (user.CheckUser())
            {
                Session["user"] = user;
//Session["Pkey"] = user.Skey;
                Response.Redirect("Home.aspx");
            }
else
            {
MsgBox.Show(this, "Invalid email id or password");
            }
        }
    }

}
```

Inserting data

```
if (!user.CheckEmailId())
        {
if (user.Insert() != 0)
            {
```

```
if (user.CheckEmailId())
                {
                    user.LoadUser();
                    ad.UId = user.Id;
                    ad.Insert();
                }
```

Storage

```
Users user;
protectedvoid Page_Load(object sender, EventArgs e)
    {
if (Session["user"] != null)
            user = (Users)Session["user"];
        Label1.Text = "";
    }
Content ct = newContent();
GContent gt = newGContent();
Group gp = newGroup();
GMember gm = newGMember();
protectedvoid ImageButton1_Click(object sender, ImageClickEventArgs e)
    {
try
        {
```

Rertival

```
if (RadioButton1.Checked)
        {
            ct.Symptoms = TextBox1.Text;
            ct.UId = user.Id;
List<string> ContentInfo = ct.SearchIdBySymptoms();
            Label1.Text = "";
if (ContentInfo.Count == 0)
            {
MsgBox.Show(this, "There is no content for you.");
            }
for (int i = 0; i < ContentInfo.Count; i++)
            {
Content ct1 = newContent();
                ct1.Id = ContentInfo[i];
                ct1.LoadInfo();
Image img = newImage();
                img.Width = 70;
                img.Height = 50;
                img.ImageUrl = "~/Img/" + ct1.Report;
                Label1 = newLabel();
```

Security

```
protectedvoid ImageButton1_Click(object sender, ImageClickEventArgs e)
    {
if (Session["Pkey"] != null)
        {
if (TextBox1.Text.Equals(Session["Pkey"].ToString()))
            {
                Response.Redirect("PSearch.aspx");
            }
        }
else
        {
MsgBox.Show(this ,"Your personalized key is invalid or expired.");
        }
    }
protectedvoid LinkButton2_Click(object sender, EventArgs e)
    {
        Response.Redirect("~/KeyServer");
    }

}
```

Important queries

```
[Ctype] [int] NULL,

 CONSTRAINT [PK_ContentTab] PRIMARY KEY CLUSTERED

(

	[Id] ASC

)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]

) ON [PRIMARY] TEXTIMAGE_ON [PRIMARY]



PRIMARY KEY CLUSTERED

(

	[Id] ASC

)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]

) ON [PRIMARY]



GO

/****** Object:  Table [dbo].[SKeyTab]    Script Date: 5/19/2021 6:40:52 PM ******/

SET ANSI_NULLS ON

GO
```

```
SET QUOTED_IDENTIFIER ON

GO

SET ANSI_PADDING ON

GO

CREATE TABLE [dbo].[SKeyTab](

	[Id] [int] IDENTITY(1,1) NOT NULL,

	[UID] [int] NULL,

	[SKey] [varchar](50) NULL,

 CONSTRAINT [PK_SKeyTab] PRIMARY KEY CLUSTERED

(

INSERT [dbo].[ContentTab] ([Id], [Symptoms], [Department], [Diagnosys], [Prescription], [Precautions],
[Report], [UId], [Ctype]) VALUES (4,
N'0XWDh5CkdJXExp81I/AFF+D9IOYswsH9u/OwS6rOfockkwEAVBSvHdu67z9uidQlMXmJctcauKl2MHrdt
LfjVISkmq+AeHeMHoJfMF/HJvQTOXlJLo5pI6AkbFeI2UI6Gz3spjrp171gBPtH2remPPyjK/AKlj0pAEofLYr
UTOJXOnzCD4rWpU6smMT23zOGx8t3QCn4zlORKq+WJoFW7uZAI6EPtGM4HCOg4q1ei7pJGNh96Wgr1
c98aJyVcCPWuF2+8RwERyUjfn7JSWtuR7yGl4e7STAUMdp6emS+nQo63eOxETk5EgD8IFD397LqK1BWU
g/nsFY9e13RHWJyewVtprwGv1qVgi2CMlulCYXj0MOrQVte9PW+ZVeXjw70jbXHB80T7xTV4+TIBD7PI/r
y4x7TPzzTIKu6SjO5HAdBQJKeHx8ML+RAba5OyScq',
N'm8+PNqG0XOtJ1ZNMtrusGhD4G+PLNmy23wihvM6C/gtrLZAoPGGWFGw/5VtyXT1X',
```