An Efficient and Privacy-Preserving Disease Risk Assessment using Logistic Regression

Dr. S Rajesh[1], R Saranya[2], M Sneha[3],

[1,2,3]Department of Information Technology, Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu.

Abstract--Disease risk assessment systems are thought to have a great deal of potential to assist future smart cities and communities with their medical treatment issues because they can extract disease risk factors from a variety of patient characteristics, provide doctors with access to diagnostic resources, and reduce the amount of time that patients must undergo medical treatment. Disease risk assessment services are still unable to grow due to substantial obstacles including data security and privacy. In this paper, we present an efficient and Privacy-Preserving Disease Risk Assessment Approach named CARAD, a practical and confidential method based on Logistic Regression for evaluating disease risk in healthcare datasets. A healthcare provider can utilize CARAD to securely train a disease risk prediction model using different healthcare data from numerous hospitals and give users (such as patients and clinicians) disease risk prediction services while maintaining their privacy. During the stages of illness risk prediction and model training, all sensitive data is handled over ciphertexts without decryption. Private information of both users and healthcare providers can be effectively protected as a result.

Index Term: Disease Risk Analysis,security,secure data training,Logistic Regression,privacy-preserving.

1.INTRODUCTION

IN RECENT YEARS, it has been noted that disease prediction systems have become more prevalent as a result of the rapid expansion of data mining techniques. Sharing of data is ongoing in order to improve disease prediction models. As brilliant as this may sound, a number of issues, such as information security and forecast effectiveness, could restrict its usage in a real-world setting. A large amount of data is already being collected in an electronic healthcare system under big data-driven society to produce insights on disease prediction and to enhance patient care[1]. Online disease risk assessment, one of the most well-liked e-healthcare applications, is revolutionising traditional medicine since it can identify a risk condition before it develops into an illness or disorder, and the cost of intervention is far lower than the final cost of treatment.[2] Model training and disease risk prediction are the two basic components of disease risk assessment in general. The e-healthcare provider gathers and aggregates local medical data from various medical centres during the model training phase, then trains a disease risk prediction model based on machine learning techniques[3] During the disease risk prediction phase, the e-healthcare provider can provide users with online disease risk prediction services using the trained prediction model, which will significantly increase the effectiveness of medical care and peoples' quality of life.

As a result of the diagnosis models being an important resource for healthcare providers, neither party is eager to share any information with unreliable parties. Therefore, the mechanisms that can be used to generate diagnosis models in a distributed environment while safeguarding the privacy of medical data are constrained by the aforementioned factors[4]. Diagnostic mistakes in medicine are typically brought on by two factors

First, when local medical data are outsourced to an e-healthcare provider, patients' private information and the clinical treatment plans of the medical centres may be disclosed. Local medical data often contain large patient treatment records and statistics data of medical centres[4]. Second, the model for predicting disease risk that has been trained is frequently seen as a valuable corporate asset. The e-healthcare provider could suffer a direct financial loss as a result of the prediction model leak. Thirdly, since users' health conditions, illnesses, and medication situations may be revealed during the disease risk prediction, users' disease risk query requests and corresponding query results are also highly sensitive[5],[6]. Unfortunately, as demonstrated by the majority of e-healthcare research studies [7]–[10], security and privacy concerns have seriously hampered

the widespread adoption of e-healthcare systems. After all, the misuse and exposure of personal health information (PHI) would result in serious privacy leakages, let alone the involvement of third-party CPs who are not entirely trusted [11], [12].

Numerous secure multi-party computation (SMC) techniques [13, 14] and homomorphic encryption [15, 16] have been proposed as privacy-preserving medical data processing methods to address the aforementioned challenges. Homomorphic encryption, in particular, allows for mathematical operations over ciphertexts,

However, the majority of homomorphic encryption-based schemes incur significant computational overhead.

In this research, we present an efficient and privacy-preserving disease risk assessment approach for healthcare datasets based on logistic regression. With the help of CARAD, the online healthcare provider may safely train a disease risk prediction model over heart disease dataset from various healthcare facilities and offer consumers disease risk prediction services while protecting their privacy. This approach allows for adequate security of the private data of users, e-healthcare providers, and medical facilities.

The three primary contributions of this study are as follows:

1. First, using healthcare data, CARAD successfully trains a disease risk prediction model. Even when medical facilities in CARAD receive many attributes from cases, the disease risk prediction model can still be successfully trained by the e-healthcare provider. Additionally, a method for updating the prediction model has been developed that enables medical centres to upload newly acquired medical data for routine model updates.
2. Second, CARAD protects user privacy when determining illness risk and training models. We suggest a modified Elgamal cryptosystem in CARAD so that the prediction model can be securely trained without revealing private information about medical facilities ,protecting user requests for and results from disease risk queries as well as the e-healthcare provider's disease risk prediction model. Therefore, CARAD protects all sensitive data.
3. Third, CARAD performs model training and illness risk prediction with computational and communication efficiency better than other cryptosystems. With data pre-processing in healthcare facilities, the encryption times and communication overhead are significantly decreased during the model training phase. Furthermore, by optimising the mathematical operations used in logistic regression, the high efficiency of the disease risk prediction phase is also guaranteed. The analysis using actual medical datasets demonstrates the effectiveness of CARAD.

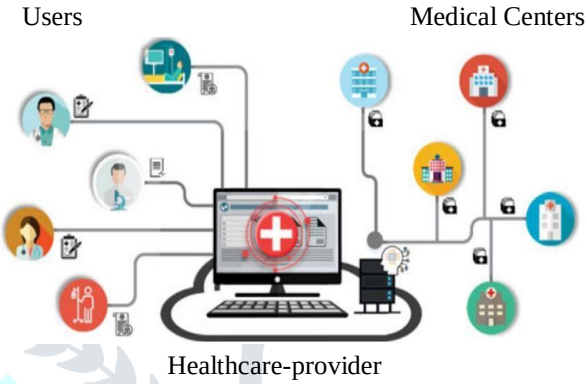

Fig:1 Architecture of online disease assessment approach by the healthcare providers.

The rest of this paper is structured as follows. We formalise the models and specify our design objective in Section 2. With Logistic regression serving as preliminary analysis, Section 3 reviews the Elgamal cryptosystem and disease risk prediction model. Then, in Section 4, we outline our CARAD. Sections 5 then present the security analysis and performance evaluation, respectively. In Section 6, we also examine a few related works. In Section 7, we finally reach our conclusions.

The following desirable qualities should, in theory, be offered by a promising disease risk prediction system.

1) ***Completeness:*** The disease risk prediction must simultaneously provide disease model training and distant disease prediction.

2) ***Privacy-Preservation***: A disease risk prediction system's success is significantly influenced by privacy-preservation [16], [17]. Naturally, confirmed patients would not want to offer their data for training if the privacy problem is not adequately addressed. Patients who had not yet received a diagnosis would not use the prediction service.

3) ***Efficiency*** is a key factor in determining whether a disease risk prediction system can be implemented in the practice are resulting in high productivity .

4) . Consider the emergency situation where a user unexpectedly faints at homeThe individual symptoms gathered by medical monitoring devices can be forwarded to the HP to acquire the disease prediction in order to perform the appropriate first-aid actions.time. As a result, the response time should be as quick as possible, taking into account both the computing cost and communication. **Overhead** must be productive.

2. MODELS, SECURITY SPECIFICATIONS, AND DESIGN OBJECTIVES

This section identifies our design objectives as well as formalises the system model and threat model utilised in this paper.

System Model

In our system model, we primarily concentrate on how to provide disease risk prediction and medical data training for online disease risk assessment systems while protecting patient privacy. Every medical facility and user has a computer or mobile device that can connect to the e-healthcare provider. The system is made up of four distinct components: users, medical centres, healthcare providers, and trusted authorities (MCs,HP,TA).

The entire system is booted up by **TA,** which is a trusted authority (i.e., a government agency), by establishing system settings and distributing keys to hospitals, e-healthcare providers, and users.

Medical centers (MCs) is a collection of m hospitals. Each MC in our system is the owner of its local medical dataset. Additionally, each MC will carry out the pre-processing and encryption procedures to produce its encrypted local training data and contract out the ciphertexts to the e-healthcare provider. Keep in mind that in our scheme, various medical centres may acquire various qualities.

Healthcare Providers is an online healthcare service that provides illness risk assessment as part of its e-healthcare services. HP is in charge of gathering the local training data that is encrypted from various medical facilities, training the disease risk prediction model, and providing users with privacy-preserving disease risk prediction services.

In the e-healthcare system, **users** are either doctors or patients, and they are denoted by the letter U. Each U_I a set of symptoms gathered by medical sensors that it can utilise to create an encrypted illness risk inquiry request, as well as to access disease risk prediction services from HP.

Security Requirements

We take into account that users, HP, and MCs are sincere but curious when developing our threat model. In particular, MCs and HP carry out actions during the model training process honestly, but an MC seeks out local training data from other MCs and the prediction model of the HP for commercial purposes. Furthermore, HP is also avaricious with regard to each MC's regional training data. Additionally, while HP and users strictly follow the protocol during the disease risk prediction process, HP makes an effort to examine the precise symptom vectors and user-generated disease risk query results. The HP's disease risk prediction model is another topic of interest for consumers. Keep in mind that an e-healthcare system may be subject to additional threats, such as poisoning and denial-of-service attacks.

These threats are currently outside the scope of this research and will be taken into consideration in subsequent work because our goal is to secure sensitive data of MCs, HP, and users in disease risk assessment. The following security standards ought to be met in light of the above security concerns.

Confidentiality: Ensuring the safety of the HP's prediction model and the local training data for MCs. Both the trained disease risk prediction model and the local training data are typically recognised as the exclusive property of individual businesses. Therefore, the local training data of MCs and the HP prediction model cannot be revealed during the disease risk assessment.

Privacy: preventing HP from accessing users' symptom vectors and disease search results. Since the results of the disease risk query and the symptom vector can reveal a user's private information, it is important to make sure that HP is kept in the dark about the precise symptom vectors and final disease risk query results of users.

Robustness: computational overhead is productive

Design Objective

Our design objective is to provide an effective and privacy-preserving online medical prediagnosis framework under the aforementioned system model and security criteria.

In particular, the three goals listed below must be met.

***Performance*:** Recognise vertical data distribution for training and updating medical information. Medical data are typically distributed among several medical facilities, and various facilities may gather various attributes from instances. In light of this,the model should be effective in handling various kinds o attributes and their instances.

Maintenance of privacy and security: The disease risk assessment system has a persistent concern with the privacy and security of medical data. If the private medical information of MCs, HPs, and users is made public, severe repercussions could result. Therefore, it is important to ensure the security of the local training data for MCs, the disease risk prediction model for EP, and the symptom vectors/disease risk query results for users.

Low communication and computation costs:Despite the fact that servers and mobile devices' processing power is growing quickly, high-bandwidth, low-delay communication between MCs and HP is available. It is still challenging for HP to manage vast amounts of medical data. Mobile device users' batteries continue to have a finite capacity. Taking into account the aforementioned considerations, the suggested approach should achieve high-efficiency in terms of computing and communication.

3 PRELIMINARIES

We examine Elgamal cryptosystem, logistic regression, and introduce the disease risk prediction with logistic regression in this section, which will provide the framework for our scheme.

Elgamal Cryptosystem

Key Generation:

1. Choose a large prime number, p, and a primitive root, g, modulo p.
2. Select a random integer, x, such that 1 <= x <= p-2.
3. Compute $y = g^x \bmod p$.
4. The public key is (p, g, y), and the private key is x.

Encryption:

1. Choose a random integer, k, such that 1 <= k <= p-2.
2. Compute a shared secret, $s = g^{(k*x)} \bmod p$.
3. Encode the plaintext message, m, as an integer in the range 0 <= m < p.
4. Compute the ciphertext as (c1, c2), where $c1 = g^k \bmod p$ and c2 = m*s mod p.
5. Send (c1, c2) to the recipient.

Decryption:

1. Compute the shared secret, $s = c1^x \bmod p$.
2. Compute the plaintext message, $m = c2 * s^{(-1)} \bmod p$, where $s^{(-1)}$ is the modular multiplicative inverse of s modulo p.
3. Decode m back into the original plaintext message.

Note that this algorithm assumes that the plaintext message is an integer in the range 0 <= m < p. If the message is a string or some other type of data, it must be first converted into an integer before encryption, and then back into its original form after decryption.

Also, it is important to choose appropriate values for p and g to ensure the security of the system. In practice, large primes (typically hundreds or thousands of bits long) are used to prevent attacks such as brute force and discrete logarithm attacks.

Logistic Regression.

Logistic regression is a popular classification algorithm that can be used to predict the probability of an instance belonging to a certain class.

Given a set of training data consisting of input vectors X and binary class labels y (0 or 1), the goal is to learn a set of parameters w such that the logistic function $g(z) = 1 / (1 + e^{(-z)})$, where $z = w^T X$, predicts the probability of y=1.

The logistic regression algorithm can be trained using gradient descent, a method for finding the optimal set of parameters w that minimizes the logistic loss function, which measures the difference between the predicted probabilities and the actual labels in the training data.

Here's the algorithm for training a logistic regression model:

1.Initialize the parameters w to small random values.

2.Compute the logistic function for each instance in the training data:

z = w^T X

p = g(z)

3. Compute the logistic loss function for the entire training data:

$J(w) = -1/m * (y^T \log(p) + (1-y)^T \log(1-p)) + \text{lambda}/2m * ||w||^2$ where m is the number of training instances, lambda is the regularization parameter, and ||w||^2 is the L2 norm of w.

4. Compute the gradient of the loss function with respect to w:

grad_J(w) = 1/m * X^T (p - y) + lambda/m * w

5.Update the parameters using gradient descent:

w = w - alpha * grad_J(w) where alpha is the learning rate.

Here's the algorithm for predicting the class of a new instance:

1. Compute the logistic function for the new instance:

$z = w^T X\_new$

p = g(z)

2. If p >= 0.5, predict the class as 1, otherwise predict the class as 0.

Disease Risk Prediction With Logistic regression For an e-healthcare system, logistic regression can be usedto forecast the likelihood of contracting various diseases.We now introduce the use of logistic regression to estimate disease risk

Additionally, disease risk can be predicted using logistic regression. Based on its input features-in this case, symptoms-the logistic regression model calculates the likelihood that a given instance belongs to a specific class (in this case, a disease). The output of the linear regression model is mapped by the logistic function, sometimes referred to as the sigmoid function, to a probability value between 0 and 1.

The logistic regression model is able to calculate the likelihood that a new instance with the symptom vector X_0 would develop each disease. To assess whether an instance is at risk of contracting the disease or not, the anticipated

likelihood of the instance contracting the disease can be compared to a threshold number (for example, 0.5).

In conclusion, the logistic regression model can be used to forecast disease risk by calculating the likelihood that a particular instance would develop a disease based on the disease's input properties (i.e., symptoms). The risk of contracting the disease can be calculated by comparing the anticipated probability to a threshold value.

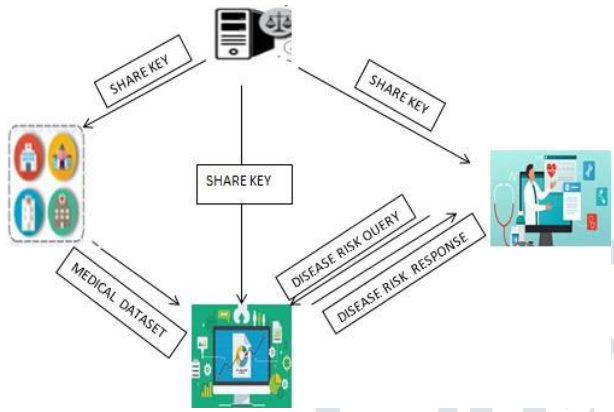

FIG:2 System model of CARAD

4 PROPOSED PRIVACY-PRESERVING SCHEME

We introduce our CARAD scheme in this part, which primarily comprises of four phases: Initialization of the system, preprocessing and encryption of medical data, safe data aggregation and training, and privacy-protecting illness risk prediction are the first three steps.Fig. 3 gives a description of CARAD's overall structure. To create encrypted local training data, each MC_i first performs data preprocessing and encryption procedures on its local medical dataset. This data is then used by EP to train the disease risk prediction model. Then, after users encrypt the symptom vectors they use to generate disease risk query requests, EP will offer the users a disease risk prediction service that protects their privacy.

4.1 System Initialization

During the system initialization phase, TA generates the system parameters and distributes keys for MCs, EP, and users using the ElGamal cryptosystem.

TA first chooses security parameters k, k1, k2, k3, k4 such that k > k1 + k2, k2 > k3 + k4, and executes Gen(k) to generate the parameters of the ElGamal cryptosystem, which includes the secret key SKp and the public key PKp. The secret key SKp consists of a large prime p and an element α in the finite field Zp*.

The public key PKp consists of the prime p and the element β = α^x mod p, where x is the secret key SKp.

Then, TA selects a large random number g satisfying |g| < k/2 and computes h = g^x mod p. Note that MCs may collect different attributes of instances, and a list L covering all collected attributes needs to be generated by TA. Finally, TA publishes the public parameters <k, k1, k2, k3, k4, p, β, h, L>.

For MCs, TA selects a random number st in Zp* as the task ID for every medical data aggregation task, and computes the secret keys $SK\_MC_i = s_t$, which are used to encrypt local training data of MCs. To encrypt the data, each MC randomly selects a secret key r_i in Zp* and computes the ciphertext (c1i, c2i) = (α^{ri} mod p, m^{ri} * plaintext mod p), where plaintext is the local training data of the MC.

For HP, TA distributes the secret key SKEP = (SKp, g) to HP, which is used to decrypt the data aggregation results. To decrypt a ciphertext (c1, c2),

HP computes the plaintext as m = c1^x * c2^(-1) mod p.

For each U_i, TA chooses two large primes p_i and a_i satisfying jp_ij = k1and ja_ij = k2, which are used for encrypting symptom vectors. Moreover, TA also sends pi to HP. To encrypt a symptom vector X = (x_1, x_2, ..., x_u), U_i randomly selects a secret key k_i in Zp_i* and computes the ciphertext (c_i, d_i) = (g^{ki} mod p_i, a_i^{ki} * x_i mod p_i).

4.2 Medical Data Preprocessing and Encryption

Every MC_i first preprocesses its gathered local medical dataset in this phase to produce local training data. Additionally, each MC_i encrypts its local training data before sending it to EP using the secret key $SKMC_i$.

Step 1. ***Medical Data Preprocessing.***

Assume that MC_i owns a local medical dataset $S^{(i)}$, which contains $l^{(i)}$ confirmed clinical instances. For each instance, MCi collects w attributes and v diseases, which can be represented as

$$X^{(i,k)} = (x_1^{(i,k)}, x_2^{(i,k)}, ..., x_w^{(i,k)}) \in \{0, 1\}^w$$
$$Y^{(i,k)} = (y_1^{(i,k)}, y_2^{(i,k)}, ..., y_v^{(i,k)}) \in \{0, 1\}^v$$

where k = 1, 2, ..., l(i).

At first, MCi checks if there are uncollected attributes in the List L. If existing, MCi pads the uncollected attributes with zero for each instance. Assume that the number of attributes in L is u, thus, X(i,k) is extended to a u-dimensional vector as

$$X^{(i,k)} = (x_1^{(i,k)}, x_2^{(i,k)}, ..., x_w^{(i,k)}) \in \{0, 1\}^w$$

Moreover, MC_i generates a vector E = (e_1, e_2, ..., e_u), where e_s = 1 if MCi collects the sth attribute in L, and e_s = 0, otherwise

.

In detail, for s = 1, 2, ..., u and t = 1, 2, ..., v, $NX^{(i)}$, $NY^{(i)}$, and $NA^{(i)}$ represent the number of instances who have symptom x_s, suffer from disease y_t, and have symptom x_s while suffering from disease y_t, respectively. Besides, in order to construct a prediction model with vertically distributed datasets, vectors $NB^{(i)}$ and $NL^{(i)}$ should be computed in our scheme. Specifically, $NB^{(i)}$ is derived from $NY^{(i)}$ and collected attributes of MC_i, while $NL^{(i)}$ is

computed with the total number of instances $L^{(i)}$ and collected attributes of MC_i.

Step 2. ***Local Training Data Encryption***

After generating local training data < $NX^{(i)}$, $NY^{(i)}$, $NA^{(i)}$, $NB^{(i)}$, $NL^{(i)}$ >, for each element $a_{(i)}$ in < $NX^{(i)}$, $NY_{(i)}$, $NA_{(i)}$, $NB_{(i)}$, $NL^{(i)}$ >,

MC_i executes encryption operations using the ElGamal cryptosystem as follows:

Generate a random number $r^{(i)}$ such that $1 \leq r^{(i)} \leq p-2$, where p is a large prime number and serves as the public parameter of the ElGam

Secure Data Aggregation and Training

Step 1.When all the encrypted local training data < $NX^{(i)}$, $NY^{(i)}$, $NA^{(i)}$, $NB^{(i)}$, $NL^{(i)}$ >, where i = 1, 2, ..., d, from all MCs are received, HP first executes aggregation operations. Specifically, for each element $a^{(i)}$ in < $NX^{(i)}$, $NY^{(i)}$, $NA^{(i)}$, $NB^{(i)}$, $NL^{(i)}$ >, i = 1, 2, ..., d, HP aggregates the ciphertexts of local training data .

After this, EP obtains the encrypted global training data < NX, NY, NA, NB, NL >:

Furthermore, for every $a^{(i)}$ in < NX, NY, NA, NB, NL >, HP decrypts it with the secret key SKEP as follows:

$a = ((L(a) \bmod N^2) * m) \bmod N$, where $L(x) = (x-1) \bmod N$

Finally, HP obtains the global training data:

$NX = (Nx_1, Nx_2, ..., Nx_W)$

$NY = (Ny_1, Ny_2, ..., Ny_v)$

$NA = (Nz_{11}, ..., Nz_{ts}, ..., Nz_{vw})$

$NB = (Nd_{11}, ..., Nd_{ts}, ..., Nd_{vw})$

$NL = (Nl_1, Nl_2, ..., Nl_w)$

Step 2. Disease Risk Prediction Model Training

With the elements in global training data < NX, NY, NA, NB, NL >, HP can train the disease risk prediction model M through logistic regression. The model predicts the probability of a patient having a disease based on their clinical and demographic data.

Let X be a matrix containing the patient data, where each row corresponds to a patient and each column corresponds to a feature. Let Y be a vector containing the labels, where each element is either 0 or 1, indicating whether the patient has the disease or not.

EP first decrypts the encrypted training data as follows:

X = Dec(NX, SKEP)

Y = Dec(NY, SKEP)

Then, HP trains the logistic regression model using X and Y:

θ = argmin(-1/m * (Y'log(sigmoid(Xθ)) + (1-Y)'log(1-sigmoid(Xθ)))), where sigmoid(z) = 1/(1+exp(-z))

Here, θ is a vector of coefficients for the logistic regression model.

Finally, EP encrypts the coefficients with ElGamal encryption and sends them to the MCs for inference:

$\theta\_enc = Enc(\theta, PK)$

The MCs can use θ_enc to perform disease risk prediction for new patients using the logistic regression model.

Algorithm 1. STDA: Secure Training Data Aggregation

Input: The local datasets of MCs: $S^{(i)}$, i = 1, 2, ..., m.

Output: The global training data: X, Y, A, B, and L.

1: MC_i preprocesses $S^{(i)}$ to generate local training data < $X^{(i)}$, $Y^{(i)}$,$A^{(i)}$, $B^{(i)}$, $L^{(i)}$ > $S^{(i)}$;

2: foreach $a^{(i)}$ in < $X^{(i)}$, $Y^{(i)}$, $A^{(i)}$, $B^{(i)}$, $L^{(i)}$ > do

3: MC_i uses its secret key SK_MC_i to encrypt $a^{(i)}$ [$a^{(i)}$ || SK_MC_i];

4: end

5: MC_i obtains the encrypted local training data < [$X^{(i)}$ || SK_MC_i], [$Y^{(i)}$ || SK_MC_i], [$A^{(i)}$ || SK_MC_i], [$B^{(i)}$ || SK_MC_i], [$L^{(i)}$ || SK_MC_i] >;

6: foreach [$a^{(i)}$ || SK_MC_i] in < [$X^{(i)}$ || SK_MC_i], [$Y^{(i)}$ || SK_MC_i], [$A^{(i)}$ || SK_MC_i], [$B^{(i)}$ || SK_MC_i], [$L^{(i)}$ || SK_MC_i] > and i = 1, 2, ..., d do

7: EP aggregates the encrypted [a(i) || SK_MCi] over ciphertexts

8: end

9: EP obtains the global training data < X, Y, A, B, L >;

Privacy-Preserving Disease Risk Prediction

This is a technical paper proposing a Privacy-Preserving Disease Risk Prediction (PDRP) algorithm that allows users to encrypt their symptom vectors and generate disease risk query requests. The algorithm then computes the disease risk query response over ciphertexts, which users can decrypt to obtain the final disease risk query results.

The proposed algorithm has three steps:

1. **Disease Risk Query Generation:** Users encrypt their symptom vectors, invert each element, and extend them to a (2u+1)-dimensional vector. They choose a large random number, select 2u+3 random numbers, and compute Cj for each element xj. The user keeps SKU_i as their secret key and sends the disease risk query request to HP.
2. **Query Response Computation**: HP receives the query request and computes the query response for each disease. HP generates two (2u+1)-dimensional vectors and computes the disease risk query responses with the query requests. For each query response, HP sets $L_{it}(2u+2) = L_{it}(2u+3) = 0$, computes D_{tj}, and then computes D_t. Finally, EP sends the query responses back to the user.

3. **Query Results Reading:** Once the user receives the query responses, they compute the query results. For each D_t, the user computes E_t and R_t. Finally, the user obtains the final disease risk query results.

The algorithm ensures privacy preservation by encrypting the symptom vectors, keeping SK_{Ui} as a secret key, and computing the disease risk query responses over ciphertexts.

Algorithm 2. PDRP with Logistic Regression and ElGamal Encryption

Input: The symptom vector X of U_i, and the disease risk prediction model M of EP.

Output: The query results R1, ..., Rv, R'1, ..., R'v.

1: U_i extends its symptom vector X with a column of ones to include the intercept term: X ← [X, 1].

2: Ui encrypts the extended symptom vector X with ElGamal encryption to obtain the ciphertext C_X.

3: Ui sets x_0 2u+2 = x_0 2u+3 = 0.

4: For each j = 1, 2, ..., 2u+3, U_i uses prime numbers a0 and p0 and random numbers cj to generate the query request ciphertexts:

Let X_j = [0, ..., 0, c_j, 0, ..., 0] be the vector with c_j in the jth position and zeros elsewhere.

U_i encrypts X_j with ElGamal encryption to obtain the ciphertext C_j_X_j.

U_i computes $C_j = (a_0^{cj} \bmod p_0, C_j\_X_j * (C\_X^{cj}) \bmod g^{p_0})$ and sends C_j to HP.

5: For each t = 1, 2, ..., v,

HP performs the following:

HP uses the logistic regression model M to compute the predicted disease risk for the i^{th} individual with symptom vector X: Dt = M(X).

HP uses the query request ciphertexts <C1, C2, ..., C2u+3> to compute the predicted disease risks for the i^{th} individual with the j_{th} feature set:

Let X_j be the vector corresponding to C_j.

HP computes D_j = M([X[:, 0], ..., X[:, j-1], X_j, X[:, j+1:], 1]) where X[:, i] denotes the ith column of X.

HP computes D_{0t} in the same way using the encrypted zero vectors.

HP encrypts the predicted disease risks with ElGamal encryption to obtain the ciphertexts C_D_t and C_D_{0t}.

EP sends C_Dt and C_D0t to Ui.

6. For each t = 1, 2, ..., v, Ui performs the following:

Ui uses its secret key SKUi to decrypt the ciphertexts C_Dt and C_D_{0t} to obtain the predicted disease risks for the tth individual with and without the jth feature:

D_t = Dec(SKU_i, C_D_t) and

D_{0t}= Dec(SKU_i, C_D_{0t}).

U_i computes the query results for the t^{th} individual with and without the jth feature using the predicted disease risks: Rt = exp(D_t) / (1 + exp(D_t)) and R_{0t} = exp(D_{0t}) / (1 + exp(D_{0t})).

7. Ui obtains the disease risk query results

<R_1, ..., R_v, R'$_1$, ..., R'$_v$>.

5. ACCURACY EVALUATION

We assess the prediction accuracy of CARAD using the BCW dataset in order to confirm its efficacy, and we compare the results to the situation without a privacy-preserving method.

In order to simulate the local medical datasets of 3 MCs, we first select 3 subsets from the BCW dataset. Particularly, MC1 owns 150 instances with attributes 3 through 9, MC2 owns 150 instances with attributes 1 through 7, and MC3 owns 183 instances with all 9 attributes. Additionally, since each attribute in the BCW dataset has a value between 1 and 10, each attribute should be expanded to a 10 dimension before the dataset is converted to binary. Then, we train two disease risk prediction models using the three local medical datasets (one is trained over plaintext, while the other is trained using CARAD) and test their predictive accuracy using the remaining 200 instances in the BCW dataset. The test results are shown in Table 2 it is clear that our suggested plan may produce very accurate disease risk prediction results, and that the accuracy is unaffected by the privacy-preserving technique.

After using machine learning algorithms on the Wisconsin Diagnostic dataset for breast cancer .As performance indicators, we employed Confusion Matrix, Accuracy, Precision, Sensitivity, F1 Score to assess and contrast the models and determine the optimal algorithm for breast cancer Prediction.

TABLE 1

INDICATOR VALUES FOR DIFFERENT MODELS

Models	KNN	SVM	Naïve Bayes	Logistic Regression
Accuracy	0.94	0.95	0.952	0.96

Accuracy: Accuracy is a performance metric that has the correct predictions for the test data. Accuracy is calculated as of correct predictions to the total no of predictions of the model.

Accuracy = TP+TN/TP+ TN + FP+ FN

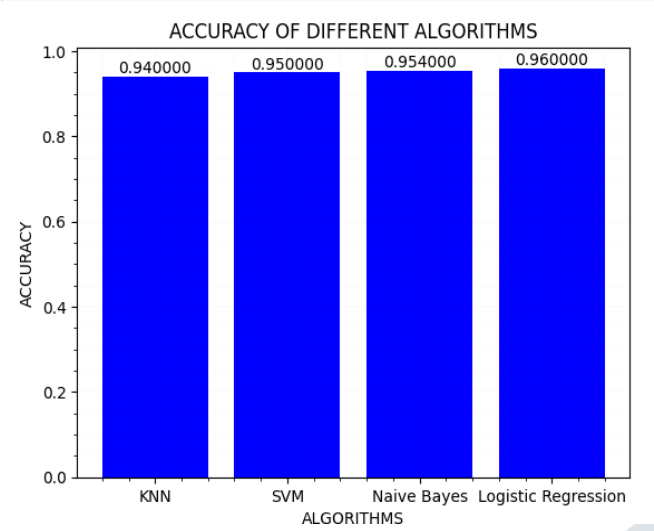

FIG:3 Accuracy evaluation on various algorithms.

In Fig 3. Among all the algorithms performed so far Logistic Regression algorithm has shown the best performance,
whereas the performance shown by other machine learning are relatively low.

A confusion matrix can be used to assess a classification problem where the outcome can be one of two or more types of classes.

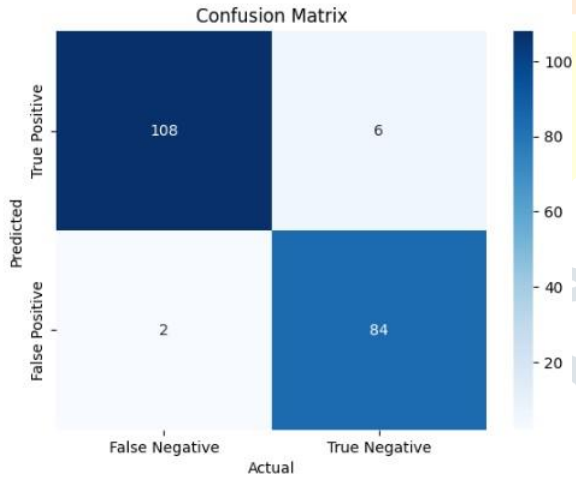

FIG:4 Confusion matrix

A confusion matrix is comprised of an actual and expected dimension, as well as
"True Positives (TP)," "True Negatives (TN)," "False Positives (FP)," and "False Negatives (FN)" in each of the two dimensions.

TABLE 2

Accuracy Evaluation of CARAD Compared with PlainText Training

	Over Plaintext	CARAD
Benign	84/86 (97.8%)	84/86 (97.8%)
Malignant	108/114 (94.7%)	108/114 (94.7%)
Overall	192/200 (96.0%)	192/200 (96.0%)

6.RELATED WORK

Liu et al. [19] introduced a new secure patient-centric clinical decision support system that retained the sensitive data in both model training and prediction stages in order to accomplish integrated privacy preservation. This system, however, was difficult to implement in practise because it called for intricate mathematical procedures. The computational complexity of mathematical issues is the foundation for the security of cryptographic techniques. Any improvement in the mathematical solutions to these issues or in processing capacity can make a cryptographic method insecure.Yang et al.[20] devised a super-increasing sequence to minimise the dimension of datasets and increase the effectiveness of their illness risk prediction scheme, which is based on naive Bayesian classification. However, it is challenging to use in high-dimensional data encryption due to the exponential growth of the coefficients of super-increasing sequences. The underlying thread across these issues is that when dimensionality rises, the size of the space expands so quickly that the amount of information becomes sparse. The amount of data required to produce a credible result frequently increases exponentially with the dimensionality.

In order to safeguard both the classifier and the user's sensitive data, Ma et al.[21] presented the privacy-preserving random forest classification over encrypted data with secure rational computing procedures for disease prediction. The integrated privacy-preserving from data training to illness risk prediction cannot be achieved by these schemes, which solely deal with the disease risk prediction phase.due to the significance of PPML in preventing data leaks in machine learning systems. To enable the collaborative training of ML models from numerous input sources without disclosing sensitive or private information, PPML uses a variety of privacy-enhancing techniques.

In order to provide patients with high-accuracy disease risk prediction services, Ayday et al. [22] combined the genomic, clinical, and environmental data and proposed a privacy-preserving disease risk prediction scheme with homomorphic encryption. However, higher the computation cost will result in higher mathematical operations and more computational time. Fully homomorphic encryption for computationally intensive applications is still economically unfeasible due to slow calculation speed or accuracy issues.

Zhou et al.[23]has put forth a brand-new safe data processing technique that permits operations on ciphertexts that are both homomorphic addition and multiplication. An effective and privacy-preserving dynamic medical text mining and picture feature extraction approach was suggested based on the proposed protocol. The majority of the aforementioned approaches only succeed in training data while protecting privacy. Additionally, the need for extensive interactions between cloud servers and data providers results in significant communication overhead in actual usage.

7. CONCLUSION:

In this paper, we have proposed an efficient and privacy – preserving disease assessment approach over medical data named CARAD. Based on Modified Elgamal cryptosystem in CARAD. Both model training over medical dataset and privacy over the disease risk prediction can be achieved effectively. Moreover, the machine learning algorithm(Logistic Regression) used here, gives better accuracy in disease risk assessment over medical dataset collected from various medical centers while ensuring data privacy and security.

.

REERENCES:

[1] Shetty, and D. Bowden, “Towards secure and smart healthcare in smart cities using blockchain,” in Proc. IEEE Int. Smart Cities Conf., 2018, pp. 1–4.

[2] J. S. Lin, C. V. Evans, E. Johnson, N. Redmond, E. L. Coppola, and N. Smith, “Nontraditional risk factors in cardiovascular disease risk assessment: Updated evidence report and systematic review for the US preventive services task force,” J. Amer. Med. Assoc., vol. 320, no. 3, pp. 281–297, 2018.

[3] L. Jena, S. Nayak, and R. Swain, “Chronic disease risk (CDR) prediction in biomedical data using machine learning approach,” in Proc. Advances Intell. Comput. Commun., 2020, pp. 232–239.

[4] C. Xu, N. Wang, L. Zhu, K. Sharif, and C. Zhang, “Achieving searchableandprivacy-preservingdatasharingforcloud-assistedE-healthcare system,” IEEE Internet Things J., vol. 6, no. 5, pp. 8345–8356, Oct.2019.

[5] H. Lin, J. Shao, C. Zhang, Y. Fang, Cam: cloud-assisted privacy preserving mobile health monitoring, IEEE Trans. Inf. Forensics Secur. 8 (6) (2013) 985–997.

[6].L. Yang, Q. Zheng, and X. Fan, “RSPP: A reliable, searchable and privacy-preserving e-healthcare system for cloud-assisted body areanetworks,”inProc.IEEEConf.Comput.Commun.,2017,p p.1–9

[7] A. Toninelli, R. Montanari, and A. Corradi, “Enabling secure service discovery in mobile healthcare enterprise networks,” IEEE Wireless Commun., vol. 16, no. 3, pp. 24–32, Jun. 2009.

[8] Y. Ren, R. Werner, N. Pazzi, and A. Boukerche, “Monitoring patients via a secure and mobile healthcare system,” IEEE Wireless Commun., vol. 17, no. 1, pp. 59–65, Feb. 2010.

[9] M. Li, W. Lou, and K. Ren, “Data security and privacy in wireless body area networks,” IEEE Wireless Commun., vol. 17, no. 1, pp. 51–58,Feb. 2010.

[10] J. Zhou, Z. Cao, X. Dong, X. Lin, and A. V. Vasilakos, “Securing m-healthcare social networks: Challenges, countermeasures and future directions,” IEEE Wireless Commun., vol. 20, no. 4, pp. 12–21, Aug. 2013.

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

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

[13] K. Y. Yigzaw and J. G. Bellika, “Evaluation of secure multi-party computation for reuse of distributed electronic health data,” in Proc. IEEE-EMBSInt.Conf.Biomed.HealthInformat.,2014,pp.219–222.

[14] D. Zhu et al., “CREDO: Efficient and privacy-preserving multilevel medical pre-diagnosis based on ML-kNN,” Inf. Sci., vol. 514, pp. 244–262, 2020.

[15] R. Bocu and C. Costache, “A homomorphic encryption-based system for securely managing personal health metrics data,” IBM J. Res. Develop., vol. 62, no. 1, pp. 1:1–1:10, 2018.

[16] X. Liu, R. Lu, J. Ma, L. Chen, and B. Qin, “Privacy-preserving patient-centric clinical decision support system

on naı ¨ve Bayesian classification," IEEE J. Biomed. Health Informat., vol. 20, no. 2, pp. 655–668, Mar. 2016.

[17] K. Y. Yigzaw and J. G. Bellika, "Evaluation of secure multi-party computation for reuse of distributed electronic health data," in Proc. IEEE-EMBSInt.Conf.Biomed.HealthInformat.,2014,pp.219–222.

[18] D. Zhu et al., "CREDO: Efficient and privacy-preserving multilevel medical pre-diagnosis based on ML-kNN," Inf. Sci., vol. 514, pp. 244–262, 2020.

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

[20] X.Yang, R. Lu, J. Shao, X. Tang, and H. Yang, "An efficient and privacy-preserving disease risk prediction scheme for e-healthcare," IEEE Internet Things J., vol. 6, no. 2, pp. 3284–3297, Apr. 2019.

[21] Z. Ma, J. Ma, Y. Miao, and X. Liu, "Privacy-preserving and highaccurate outsourced disease predictor on random forest," Inf. Sci., vol. 496, pp. 225–241, 2019.

[22] E. Ayday, J. L. Raisaro, P. J. McLaren, J. Fellay, and J. Hubaux, "Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data," in Proc. USENIX Workshop Health Inf. Technol., 2013, Art. no. 1

[23]J. Zhou, Z. Cao, X. Dong, and X. Lin, "PPDM: A privacy-preserving protocol for cloud-assisted e-healthcare systems," IEEE J. Sel. Topics Signal Process., vol. 9, no. 7, pp. 1332–1344, Oct. 2015.