J. Inst. Eng. India Ser. B (June 2022) 103(3):817–830
https://doi.org/10.1007/s40031-021-00696-1

ORIGINAL CONTRIBUTION

ElGamal Homomorphic Encryption-Based Privacy Preserving Association Rule Mining on Horizontally Partitioned Healthcare Data

Nikunj Domadiya[1] · **Udai Pratap Rao**[2]

Received: 26 August 2020 / Accepted: 20 October 2021 / Published online: 4 January 2022

Abstract In today's world, life-threatening diseases have become a pre-eminent issue in healthcare due to the higher mortality rate. It is possible to lower this mortality rate by utilizing healthcare intelligence to detect diseases early. Patient's medical data is stored in the EHR system, which is kept up to date by the healthcare provider. Data mining techniques like Association Rule Mining can detect a patient's disease from their symptoms using digital healthcare data stored in the EHR system. Association rule mining's efficacy can be improved by using global data from various EHR systems. It mandates that all EHR systems exchange healthcare records to a central server. When personal health information is made available on an untrusted server, several privacy laws may be violated. As a result, the challenge of privacy preserving distributed healthcare data mining has become a well-known study field in the healthcare industry. This research uses an efficient ElGamal homomorphic encryption technique to protect privacy in a distributed association rule mining. The proposed approach to discover the risk factor of most life-threatening diseases like breast cancer and heart disease with its symptoms and discuss the scope for combating COVID-19. Theoretical analysis of the proposed approach shows that it is efficient and maintains privacy in an insecure communication environment. An experimental study with a real dataset shows the proposed approach's benefit compared to the local single EHR system results.

✉ Nikunj Domadiya
domadiyanikunj002@gmail.com

[1] Computer Engineering Department, L. D. College of Engineering, Ahmedabad, India

[2] Computer Engineering Department, National Institute of Technology, Surat, India

Keywords Association Rule Mining · Breast Cancer Disease · Coronavirus(COVID-19) · Data Mining Privacy · Distributed Healthcare Data Mining

Introduction

Human life-threatening diseases are the primary focus of medical research all around the world [1]. Health researchers have recently focused a significant deal of attention on COVID-19, as well as cancer and other life-threatening diseases. According to the 2015 National Vital Statistics Report (NVS) [2], cancer and heart disease are the two most common causes of death. Fatality rate from cancer and heart disease accounted for 45.3% of all U.S. deaths in 2010, according to the Department of Health and Human Services (Fig. 1). As the most deadly disease among women, breast cancer claims millions of lives each year in the USA. Figure 1 displays the number of cancer cases in the USA in 2018 for each of the major kinds of cancer [3]. As of May-2020, there have been 4,527,815 instances of Coronavirus disease (COVID-19), a rare disease that arose in 2019. Of those cases, 303,438 people have died [4]. Given the high mortality rate of these life-threatening disorders, early disease detection through an examination of the patient's symptoms is crucial to saving more lives.

Appropriate treatment and recovery of these life-threatening illnesses require early identification of the disease. Diagnostic methods for cancer and heart disease are expensive, prone to mistake, and time-consuming [5–7]. Traditionally, disease prediction relied on physician expertise rather than symptoms patterns hidden in healthcare data [8–15]. As a result, this may result in an

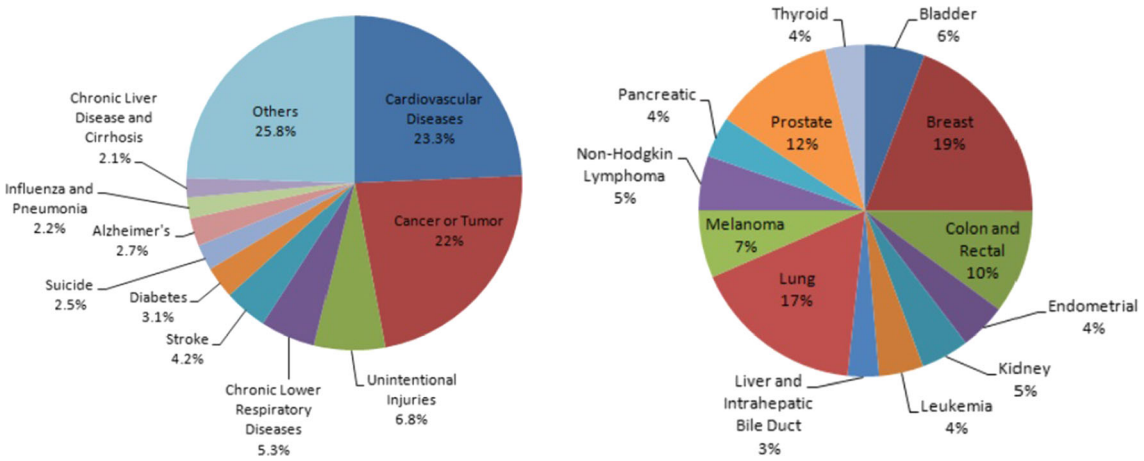

Fig. 1 Health Statistics report of USA [3]

inaccurate health diagnosis, leading to inappropriate medical treatment, which raises healthcare costs by decreasing the quality of healthcare services provided to patients [16]. Electronic healthcare record (EHR) systems are utilised in large hospitals to keep digital records. It maintains a massive amount of information on patients [17]. Data acquired in hospitals can be utilised using data mining for healthcare research and to improve healthcare services. Association rule mining is a well-known data mining approach for determining disease and symptom co-relationships [18–24]. Numerous applications of association rule mining in the healthcare area include forecasting disease based on a patient's symptoms, determining an adequate treatment for diseases, detecting medication response, and improving medical fraud detection via data mining [19, 25–29]. Association rule mining generates *IF-THEN* rules that medical professionals quickly understand. As a result, this approach is well-known amongst medical researchers and doctors for identifying the state of a disease or the appropriate treatment depending on the symptoms of the patient. As an outcome, the healthcare system becomes much more efficient in terms of cost and treatment [30].

Earlier, association rule mining on healthcare data could only be done on the EHR system of a single hospital [19, 31]. Only a limited number of patient records could be stored in a single electronic medical record (EMR). So association rule mining on the data of a single EHR system has less accuracy. Dangerous diseases (e.g. cancer and heart disease) demand more precise association rules [19, 25]. Accuracy/confidence in association rule mining can be increased by combining all EHR systems data at a central server. Patients' data must be kept private in the local EHR system since there is a threat to privacy in healthcare [32, 33]. For accurate data mining, various EHR systems must share their data while protecting privacy. As a result, medical researchers have concentrated on

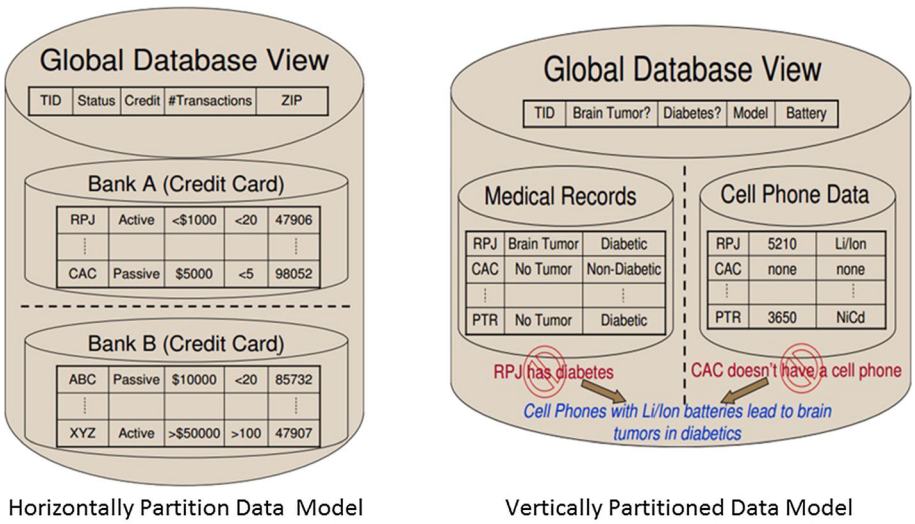

Fig. 2 Distributed Data Partition Model [49]

association rule mining on distributed healthcare data that preserves privacy. As demonstrated in Fig. 2, distributed data is either vertically or horizontally partitioned. Most large hospitals use the same EHR system schema because they follow the same standards for patient information storage in hospitals. As a result, in our study, we have included data that has been horizontally partitioned among the collaborative EHR systems [34–37]. With this insight, we're working to acquire global association rules while also safeguarding the privacy of EHR systems worldwide. UCI repository data on breast cancer and heart disease are utilised in a proposed approach for evaluating symptoms associated with both of these lives threatening diseases [38, 39].

Association Rule Mining in Distributed Data

In horizontally partitioned data, association rule mining [54] can be described as follows: Let, global healthcare dataset EHR is such a way that $EHR = \{EHR_1 \cup EHR_2 \cup EHR_3 \ldots EHR_n\}, EHR_i cap EHR_j = \emptyset, 1 \leq i \neq j \leq n$. Association rule is presented as $D \rightarrow E$, where $D \subset I$, $E \subset I$ and $D \cap E = \emptyset$, where I represents the set of itemset in dataset. The *support* and *confidence* of association rule represent the usefulness and interest measure. Algorithm 1 illustrates the association rule mining [54]. A privacy violation is occurred when the EHR system discloses a private variable (support count) in step 5 of Algorithm 1. The proposed approach computes this global support count while preserving the privacy.

Algorithm 1 Distributed Association Rule Mining

1: $L_1 = find\ frequent\_1\text{-}\ itemset(D)$
2: $for\ (k = 2;\ L_{k-1} \neq \emptyset;\ k{+}{+})\ \{$
3: $C_k = apriori\_\ gen(L_{k-1});$
4: $for\ each\ candidate\ c \in C_k$
5: $collaboratively\ find\ the\ c.count\ ;$
6: $L_k = \{c \in C_k\ |c.count \geq\ min\_sup\}$
7: $\}$
8: $return\ L = \cup_k L_k;$

Background and Related Concepts

Distributed Healthcare Data

Horizontally partitioned and vertically partitioned healthcare data are the two types of distribution of healthcare data among EHR systems.

In horizontally partitioned healthcare data, all EHR systems have an equivalent schema, but store the records of different patients. Figure 2 shows the horizontally partitioned data scenario among multiple participants. Some existing schemes for horizontal partition data in privacy preserving association rule mining are described in the literature [40–48].

Vertically partitioned healthcare data consists of many EHR systems, each with its different schema, yet they all share the same patient data. Privacy preserving association rule mining on vertically partitioned data is presented in the researches [25, 49–53]. Figure 2 shows the vertically partitioned healthcare data among hospital EHR system and the cellphone company based on common ID. Association rules discover the co-relation among diseases and cellphone usage from this vertically partitioned data.

The focus in this research is on horizontally partitioned healthcare data because most healthcare providers adopt the EHR system with the same schema.

Association rule (DE$rightarrow$F) can be derived from horizontally partitioned data by applying the following equations.

$$Support(DEF) = \frac{\sum_{i=1}^{NoofEHR} support\_count\_DEF(i)}{\sum_{i=1}^{NoofEHR} database\_size(i)} \quad (1)$$

$$Support(DE) = \frac{\sum_{i=1}^{NoofEHR} support\_count\_DE(i)}{\sum_{i=1}^{NoofEHR} database\_size(i)} \quad (2)$$

$$Confidence(DE \rightarrow F) = \frac{support(DEF)}{support(DE)} \quad (3)$$

Preliminaries

Discrete Logarithmic Problem

The DLP (discrete logarithmic problem) is described as follows :

Let $g, a \in G$, find integer p such that $g^p = a$. where g is a generator of group G, a multiplicative group of order q and, $G = < g >$.

For a large prime order group G, solving the DLP problem is proved as hard. The security of many cryptography systems depends on the assumption that solving the DLP problem is hard on such a large prime order group.

Computational Diffie–Hellman (CDH) Problem

CDH problem can be defined as follows: Given $g, g^p, g^q \in G$ where $p, q \in Z$, Without the value of p and q, determining gab is a computationally hard task, where g is a multiplicative group generator. The ElGamal cryptosystem's security holds on this CDH problem's hardness [55].

ElGamal Cryptosystem

ElGamal cryptosystem [55] is a universally known public-key cryptography system whose security is holding on the assumption of the hardness of solving CHD and DLP problems. The details of the cryptosystem are described as follows:

- *Initialization of Public Parameter* : Two large prime number a and b are selected such that $b/a-1$. Multiplicative group G of order b with generator g is selected. g, a, b are publically known parameters.
- *Public/private Keys* : Value of $p \in Z_q^*$ is selected randomly and stored as private key. Public key is calculated as $u = g^p$.
- *Encryption*: Public key u and randomly selected value $r \in Z_q$ are used for encryption of message $m \in G$ as $c = g^r$ and $d = m \cdot u^r$. The tuple $E_u(m) = (c, d)$ is a cipher text of message m using public key u.
- *Decryption* : Decryption of cipher text $E_u(m)$ using the private key p is computed as $m = d \cdot c^{-p}$.

Additive Homomorphic ElGamal Cryptosystem

The ElGamal cryptosystem can be used as additive homomorphic under group operation with the integer addition modulo q [56]. Let message m_1 and m_2 are encrypted using ElGamal encryption technique as $E_u(m_1) = (c_1, d_1) = (g^{r_1}, m_1 \cdot u^{r_1})$ and $E_u(m_2) = (c_2, d_2) = (g^{r_2}, m_2 \cdot u^{r_2})$ respectively. For additive homomorphic computation, message $m \in Z_q$ is encrypted using ElGamal encryption of g^m as $E_u(g^m)$. $E_u(g^m)$ is decrypted as g^m. Hence, to obtain the m from g^m, one additional discrete logarithm computation is required. When m is small or known to be within small range, then this computation can be implemented efficiently. Two exponentiation computations (g^r, u^r) are required for ElGamal encryption, which are independent of the message (m) and can be performed ahead of time to enhance the performance at each EHR system.

For private value m_1 and m_2, computation of $m_1 + m_2$ while maintaining privacy of individual private value can be done as follows:

1. Encrypt g^{m_1} and g^{m_2} as cipher text $E_u(g^{m_1})$ and $E_u(g^{m_2})$. Adversary cannot able to extract private value from the cipher text without private key.
2. Next, computation of $m_1 + m_2$ from $E_u(g^{m_1})$ and $E_u(g^{m_2})$ can be done as follows: $E_u(g^{m_1}) \cdot E_u(g^{m_2}) = (g^{r_1}, g_1^m \cdot u^{r_1}) \cdot \quad (g^{r_2}, g_2^m \cdot u^{r_2}) = (g^{r_1+r_2}, g^{m_1+m_2} \cdot u^{r_1+r_2}) = E_u(g^{m_1+m_2})$
3. Decryption using private key only extract value of $m_1 + m_2$ and maintain the privacy of individual value.

$$E_u(g^{m_1}) \cdot E_u(g^{m_2}) = E_u(g^{m_1} \cdot g^{m_2}) = E_u(g^{m_1+m_2})$$

Privacy Preserving Distributed Association Rule Mining for Healthcare Research

System Model

The proposed system model is organized into three distinct entities, which are as follows: (1) Key distribution server (KDS), (2) Central Server and (3) n participants or hospitals with EHR systems. The n EHR systems, which are involved in collaborative association rule mining by disclosing the healthcare data from the local EHR system to the central server. The central server authorizes them to access the data mining results of aggregated records. The central server computes association rules on collected records, which increase the effectiveness of healthcare data mining at all hospitals or assist the medical researchers in identifying the relationship between some specific diseases and it's symptoms. Figure 3 shows the system model of our approach. Each entity is described in detail as follows:

- *Key distribution server* The responsibilities of the key distribution centre are to compute and allocate keys to the central server and EHR systems.

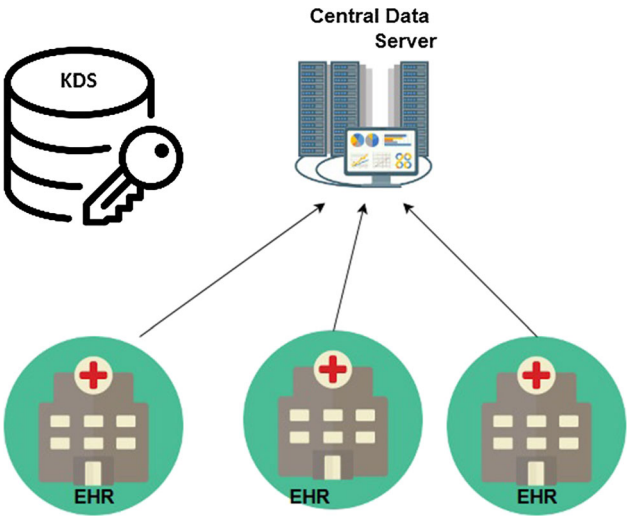

Fig. 3 System model of our scheme

– *Participant or EHR system* In the system, total n participants named as $\{EHR_1, EHR_2, EHR_3, ..., EHR_n\}$ stores the medical information of respective hospitals. Data among EHR systems are horizontally partitioned as schema at all EHR systems are the same but store different patients' information. All participating hospitals can access the result of association rule mining on aggregate data via a central server.
– *Central server* It computes the association rules on aggregated data which helps the medical researchers to discover the new relations among diseases and patient's symptoms. It shares these results to all EHR systems for improving the healthcare services at their local hospitals.

Security Model

The goal is to protect the privacy of EHR systems as it shares the sensitive/private data with the central server. This sensitive/private data should not be disclosed by the central server, any other EHR systems, or adversaries, which is not part of the collaboration.

– *Key Distribution Center (KDS)* It is assumed that the key distribution center is trusted and does not collude with any EHR systems or central server. The communication environment among EHR systems and key distribution center is assumed as secure.
– *EHR Systems or Participants* EHR systems are assumed as semi-honest, which follow the algorithm steps and give the correct input data but curious to discover other EHR system's private /sensitive data. Some EHR systems can act maliciously and cooperate with other systems to find a targeted EHR system's private key or sensitive data.
– *Central Server* Central server is assumed as semi-honest. It may collude with other EHR systems to discover the private key or data of the targeted EHR system.

Proposed Approach

It has been considered n participants (hospitals with EHR system) $\{EHR_1, EHR_2, EHR_3, ..., EHR_n\}$ with horizontally partitioned data and central server for global computation of collaborative medical research. Each $EHR_i, i \in \{1, 2, 3..., n\}$ system send its local support count of itemset I as $Icount_{(i)}$ to central server in encrypted form. Central aggregates the received data and computes the global count of itemset I as $Icount = \sum_{i=1}^{n} Icount_{(i)}$ without revealing the $Icount_{(i)}$. It is assumed that each EHR system has the computation power to compute basic ElGamal encryption operations. Key distribution center share the public parameters $< p, q, g >$ of ElGamal cryptosystem with all EHR system and central server. Each $EHR_i, i \in \{1, 2, 3..., n\}$ system also receives the private key x_i and public key $y_i = g^{x_i}$ with certificate $Cert_i$ lining with the identifier of EHR_i system from key distribution center.

All EHR system first executes the key set-up algorithm as shown in algorithm 2.

Algorithm 2 System Key set-up

Begin

1. Each $EHR_i, i \in \{1, 2, 3..., n\}$ send y_i and $Cert_i$ to central server.
2. Central server authenticate the $Cert_i$. After authentication of $Cert_i$, it send y_i and $Cert_i$ to all other EHR systems
3. After authentication of central server, each EHR system computes the global public key as $y = \prod_{i=1}^{n} y_i$

End

Communication overhead of algorithm 2 can be reduced by broadcasting all public keys and certificates of EHR systems from the central server. In our approach, only one-time execution of a system key set-up algorithm is required. Next, algorithm 3 used for computing global count of itemset I as $Icount = \sum_{i=1}^{n} Icount_{(i)}$ with privacy preserving of each EHR system.

Algorithm 3 Proposed Approach for PPDARM

Begin

1. Each $EHR_i, i \in \{1, 2, 3..., n\}$ has support count value of an itemset I as $Icount_{(i)}$.
2. Each EHR_i select a random number $z_i \in Z_q^*$ and generates a cipher text as

$$C_i = E_y(g^{Icount_{(i)}+z_i}) = (c_i, d_i)$$

and sent to central sever.

3. central server receives cipher text from all EHR systems and computes

$$C = (\prod_{i=1}^{n} c_i, \prod_{i=1}^{n} d_i) = (c, d)$$

central server sends c to all EHR system.

4. Each EHR_i computes $T_i = c^{Icount_{(i)}} \cdot g^{z_i}$ and send it to central server.
5. Central server computes $D = d \cdot (\prod_{i=1}^{n} T_i)^{-1}$
6. At last, central server computes the $log_g D$ and gets the global count as

$$Icount = \sum_{i=1}^{n} Icount_{(i)}$$

.

End

Algorithm 3 executes for all itemset's global count. In the proposed approach, MFI (maximum frequent itemset)[57] can be utilized to decrease the overall computation and communication costs.

Correctness Analysis

Central server gets the correct global count of an itemset I as $Icount = \sum_{i=1}^{n} Icount_{(i)}$ if all EHR systems honestly follow the above algorithms. At first, each EHR_i encrypt its private count value $Icount_{(i)}$ of an itemset I using random value z_i as shown in equation 4.

$$C_i = E_y(g^{Icount_{(i)}+z_i}) = (g^{r_i}, g^{Icount_{(i)}+z_i} \cdot y^{r_i}) \tag{4}$$

Then, central server aggregates all received C_i as shown in the following equation 5.

$$\begin{aligned} C &= (\prod_{i=1}^{n} g^{r_i}, \prod_{i=1}^{n} g^{Icount_{(i)}+z_i} \cdot y^{r_i}) \\ &= (g^{\sum_{i=1}^{n} r_i}, g^{\sum_{i=1}^{n} Icount_{(i)}+\sum_{i=1}^{n} z_i} \cdot y^{\sum_{i=1}^{n} r_i}) \\ &= (g^r, g^{Icount+z} \cdot y^r) = (c, d) \end{aligned} \tag{5}$$

where $Icount = \sum_{i=1}^{n} Icount_{(i)}$, $r = \sum_{i=1}^{n} r_i$ and $z = \sum_{i=1}^{n} z_i$ Next, each EHR_i receives the c and computes T_i as shown in equation 6.

$$T_i = c^{x_i} \cdot g^{z_i} = (g^r)^{x_i} \cdot g^{z_i} = (g^{x_i})^r \cdot g^{z_i} = y_i^r \cdot g^{z_i} \tag{6}$$

At last, central server computes $Icount$ using equation 7.

$$\begin{aligned} D &= d \cdot \left(\prod_{i=1}^{n} T_i\right)^{-1} = \frac{g^{Icount+z} \cdot y^r}{\prod_{i=1}^{n} (y_i^r \cdot g^{z_i})} = \frac{g^{Icount+z} \cdot y^r}{(\prod_{i=1}^{n} (y_i^r) \cdot g^z} \\ &= \frac{g^{Icount+z} \cdot y^r}{y^r \cdot g^z} = g^{Icount} \end{aligned} \tag{7}$$

Central server gets the itemset I's global support count as $log_g D$, while the local support count value remains private to the particular EHR system.

Theoretical Analysis

The communication and computing complexity of both the proposed approach and the existing approaches are evaluated.

Algorithm 2 is required to execute only once in our approach. Hence, the major cost is of algorithm-3 as it executes for computing the global count of all itemsets. *step-2* and *step-4* are executes at all n EHR system for computing the ciphertext with cost of $O(n)$. Central server processes the all received data from all EHR systems at *step-3* and *step-5* of algorithm 3 with cost of $O(n)$. Hence, the overall computation cost is $O(n)$ for a single itemset's global count computation. In terms of Communication cost of algorithm 3, all n EHR systems send the encrypted count to central server at *step-2* cause the cost of $O(n)$. Central server processes the received data and send the c to all EHR system cause the cost of $O(n)$. In *step-3*, EHR_i send T_i to central server cause the cost of $O(n)$. Next, the central

Table 1 Theoretical Analysis of Proposed Approach and Existing Approaches

	Domadiya et al. [45]	Chahar et al. [58]	Y. Jin et al. [59]	Proposed Approach
Communication Cost	O(n^2)	O(n)	O(n)	O(n)
Secure against collision among collaborative participant	No	No	Yes	Yes
Secure against collision with external adversary	Yes	Yes	Yes	Yes
Require *Trusted Third Party* for computation	No	No	Yes (*Host Computer*)	No

server computes the global count and does not require communication with EHR systems. Hence, the overall communication cost is $O(n)$. The theoretical comparison of the proposed technique with existing approaches is shown in Table 1.

Experimental Analysis

It has been described the symptoms associated with both breast cancer and heart disease in this work, along with the dataset description. Experimental analysis is performed on a system with an 4GB RAM, Intel Core i5 2.1GHz CPU and NetBeans. It has been replicated all of the records in order to reach the 80K-record total. All patient records are randomly distributed among all EHR systems, which are part of a collaboration.

Breast Cancer Analysis

Among the most deadly cancers, breast cancer is one of the most prominent. The Wisconsin breast cancer dataset [38], available freely at the UCI repository, is used for experimental analysis.

Details of Wisconsin Breast Cancer Dataset

The Wisconsin breast cancer dataset includes 32 various attributes [38]. In this study, 10 important associated attributes are taken into account out of a total of 32. The breast cancer state is denoted by the class labels *benign* and *malignant*. All attributes are in the range of 1–10. Table 2 lists the attributes of the Wisconsin breast cancer dataset.

Experiment Results Analysis

In the RHS of association rules, *class= benign or malignant* is taken into account for accuracy analysis from the experiment's results. The result of the traditional local data mining at each EHR system and the proposed approach is demonstrated in Table 3. According to the results of the experiments, the proposed approach has higher accuracy in identifying breast cancer disease than each EHR system.

The association rule's confidence value indicates the precision with which it can predict RHS from LHS. Higher accuracy at the central server is seen in Table 3 compared to individual EHR system. Using the proposed approach, medical doctors and researchers will have access to global results that will facilitate the growth of healthcare services.

Experimental analysis was performed using two different setups (1)increasing the number of EHR systems (4, 6 and 8 EHR systems) and (2) increasing the no of records in the dataset (40K, 60K and 80K records). Figures 4 and 5 show the comparison of the proposed approach and existing approaches([45, 58, 59]) in terms of computation cost while scaling of EHR systems and records. Figures 6 and 7 demonstrate the comparison between the proposed

Table 2 Attribute detail of Wisconsin Breast Cancer Dataset

Sr No	Attribute	Range
1	Clump Thickness	1–10
2	Uniformity of Cell Size	1–10
3	Uniformity of Cell Shape	1–10
4	Marginal Adhesion	1–10
5	Single Epithelial Cell Size	1–10
6	Bare Nuclei	1–10
7	Bland Chromatin	1–10
8	Normal Nucleoli	1–10
9	Mitoses	1–10
10	Class	2 for benign, 4 for malignant

Table 3 Experimental results for Breast Cancer Dataset

Association Rules	Accuracy/Confidence(%)				
	EHR1 (%)	EHR2 (%)	EHR3 (%)	EHR4 (%)	Central Data Mining Server (%)
{Bare nuclei=7 *and* Single epithelial cell size=1 *and* Normal nucleoli=2 } → {class=benign}	91	93	94	93	98
{Uniformity of cell shape=3 *and* Single epithelial cell size=2 *and* Bare nuclei=1 } → {class=benign}	93	90	91	93	99
{Single epithelial cell size=2 *and* Marginal Adhesion=1} → {class=benign}	92	91	91	90	98
{Marginal adhesion=10 *and* Uniformity of cell size=8 *and* Uniformity of cell shape=10} → {class=malignant}	94	93	92	94	99
{Mitoses=4 *and* uniformity of Cell Size=1 *and* Clump thickness=8 } → {class=malignant}	92	90	95	90	97
{Uniformity of cell shape=7 *and* Marginal Uniformity of cell size=10 *and* Bare nuclei=8 *and* adhesion=10 } → {class=malignant}	89	92	91	94	96

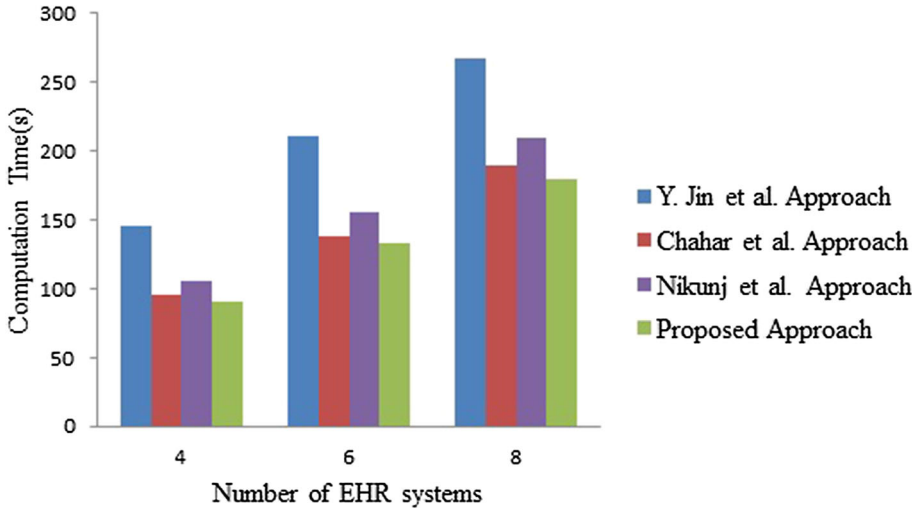

Fig. 4 Comparison of Computation Cost with Existing Approaches with Increasing Number of EHR System for Breast Cancer Dataset

approach and the existing approaches in terms of communication cost while scaling of EHR systems and records.

Analysing Heart Disease

It was also analysed the proposed algorithm using the heart disease dataset [39]. Total of 76 different attributes [39] are considered in this dataset. In this experiment, a total of 14 attributes used that are related to heart disease [60]. RHS is used as the *Class = Sick or Healthy* to choose association rules. The confidence value of these association rules from the central data mining server and each EHR system is shown in Table 4. These rules with high accuracy will help in predicting heart disease at an early stage.

Figures 8 and 9 show the comparison of the proposed approach and existing approaches ([45, 58, 59]) in terms of computation cost while scaling of EHR systems and records. Figures 10 and 11 demonstrate the comparison between the proposed approach and the existing approaches in terms of communication cost while scaling of EHR systems and records.

This experimental study demonstrates that the proposed approach outperforms existing approaches in terms of effectiveness.

Approach Proposed to Combat the Coronavirus

The new strain, which affects tens of thousands of people worldwide, is known as Novel Coronavirus, commonly known as CoViD-19 [61, 62]. It were widely undefined before the epidemic in December 2019 from its beginnings in Wuhan Province, China. As a result of the year it was created, it was given the name COVID-19.

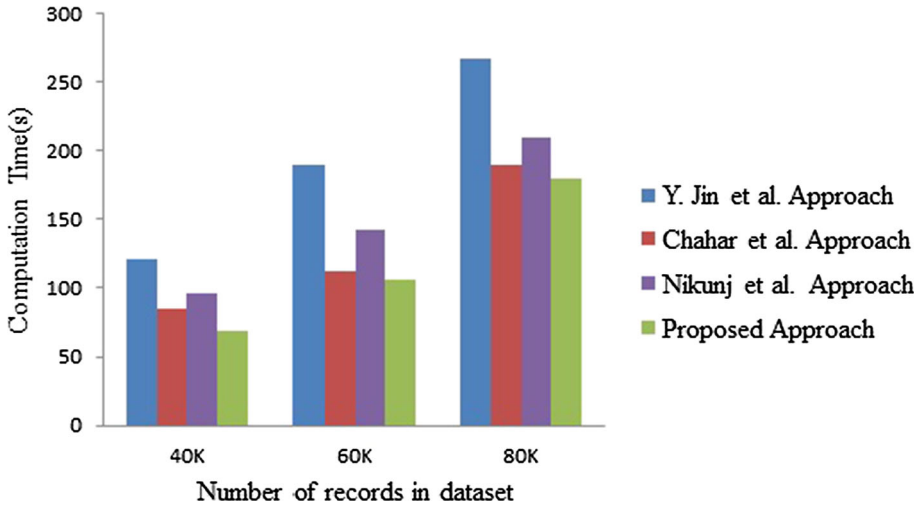

Fig. 5 Comparison of Computation Cost with Existing Approaches with Increasing Number of records for Breast Cancer Dataset

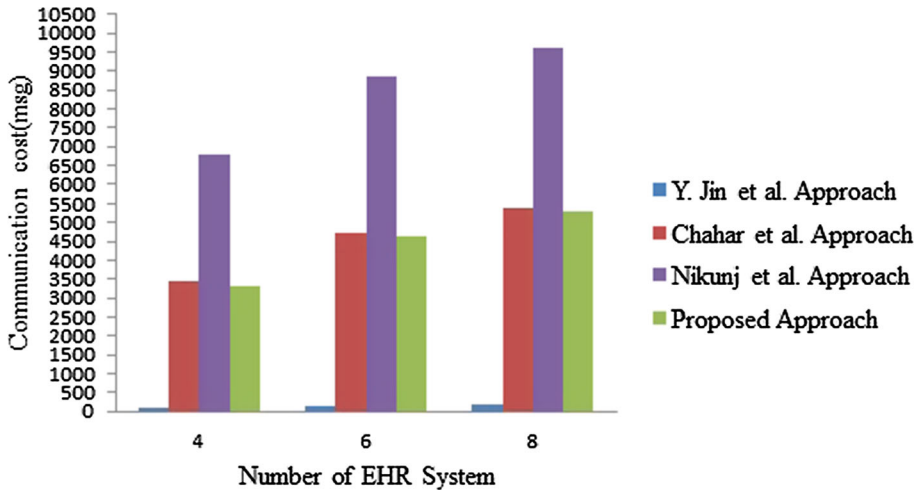

Fig. 6 Comparison of Communication Cost with Existing Approaches with Increasing Number of EHR System for Breast Cancer Dataset

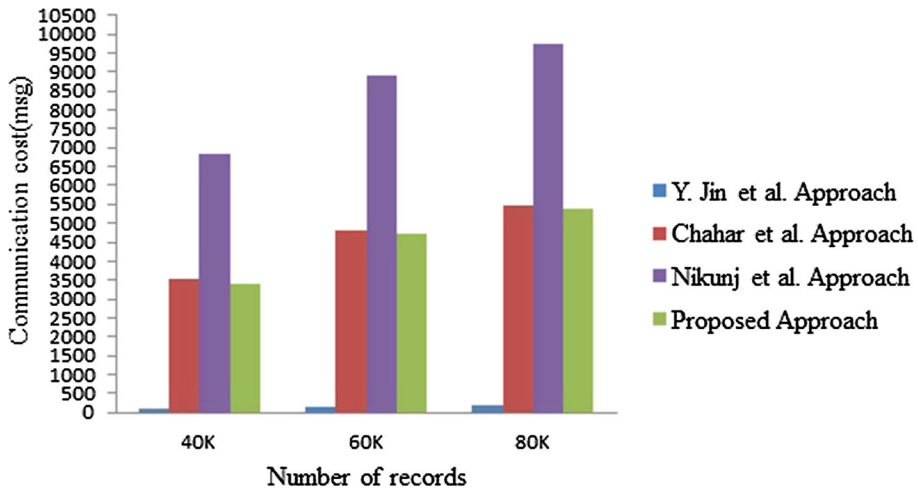

Fig. 7 Comparison of Communication Cost with Existing Approaches with Increasing Number of records for Breast Cancer Dataset

The most frequent symptoms of this disease are fatigue, fever and a dry cough. Some of the additional symptoms include headaches, nasal congestion, conjunctivitis, loss of smell or taste, and diarrhoea. Minor symptoms were observed in some Covid-19 patients. [63]. Within 14 days of coronavirus infection, the patient may experience these

Table 4 Experimental results for Heart Disease

Association Rules	Accuracy/Confidence(%)				
	EHR1 (%)	EHR2 (%)	EHR3 (%)	EHR4 (%)	Central Data Mining Server (%)
{Restecg = normal *and* Type of chest pain = asymptomatic *and* Slope = flat *and* Exercise induced angina = yes *and* Sex=female } → {Class = sick }	91	88	92	91	99
{Sex=female *and* Number of number of vessels colored = 0 *and* Exercise induced angina=no} → { Class = healthy}	92	92	91	92	98
{Type of chest pain = asymptomatic *and* Slope = flat *and* Thal = reversible defect } → { Class = sick }	92	91	91	94	96
{Thal = reversible defect *and* Exercise induced angina = yes *and* Type of chest pain = asymptomatic} → { Class = sick }	84	88	78	85	94
{Thal = normal *and* Number of vessels colored = 0 *and* Slope = up *and* Sex = male } → { Class = healthy }	85	81	75	90	93
{Sex = male *and* Type of chest pain = asymptomatic *and* Exercise induced angina = yes *and* Restecg = hypertrophy } → { Class = sick }	92	89	90	92	94
{Sex = male *and* Fasting blood sugar = no *and* Exercise induced angina = yes *and* Thal = reversible defect *and* Type of chest pain = asymptomatic } → { Class = sick }	89	91	89	89	92

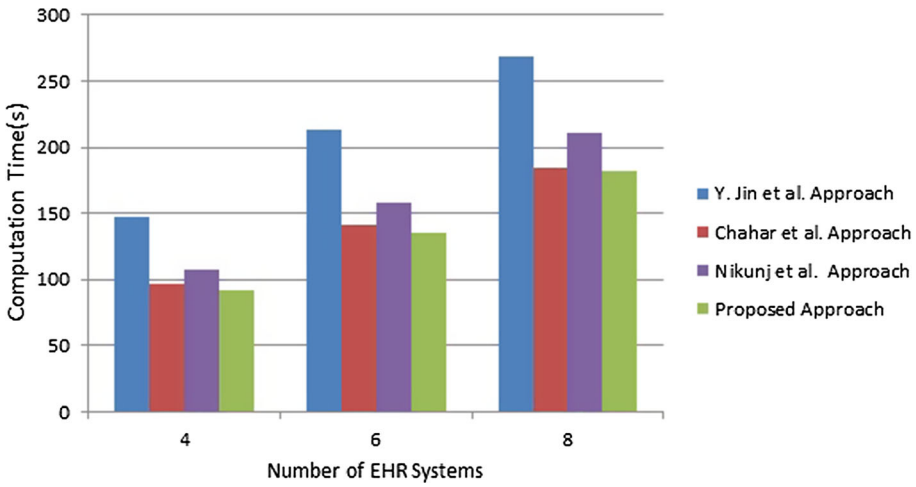

Fig. 8 Comparison of Computation Cost with Existing Approaches with Increasing Number of EHR System for Heart Disease Dataset

symptoms. One of the most difficult issues facing countries that a coronavirus has hit is the mortality rate and the time it takes to detect illness. Detecting infection early based on symptoms and selecting the right treatment to restore health to patients after infection is critical for reducing mortality and improving recovery rates from coronaviruses [61, 64–68].

Through collaborative healthcare data mining incorporating coronavirus patient data, the proposed approach can help solve the aforementioned challenges. Hospitals with coronavirus treatment facilities are using an EHR system to gather patient data, including symptoms, age, place of treatment and gender, and information on the treatment. More patient data is required for higher accuracy in data mining solutions. As illustrated in Figure 3, data mining can be used to analyse data from every hospital that adopts an EHR system and cooperates with the project.

To combat COVID-19, healthcare organizations can use global data mining technologies that incorporate data from all healthcare organizations while also safeguarding healthcare data privacy. The results of this worldwide healthcare mining can be used by doctors and medical researchers to tackle COVID-19.

Privacy Analysis

Any malicious adversary can cache all message exchange among EHR systems and reveal the private value of the targeted EHR system as a communication channel among EHR systems is insecure in our approach. The following is

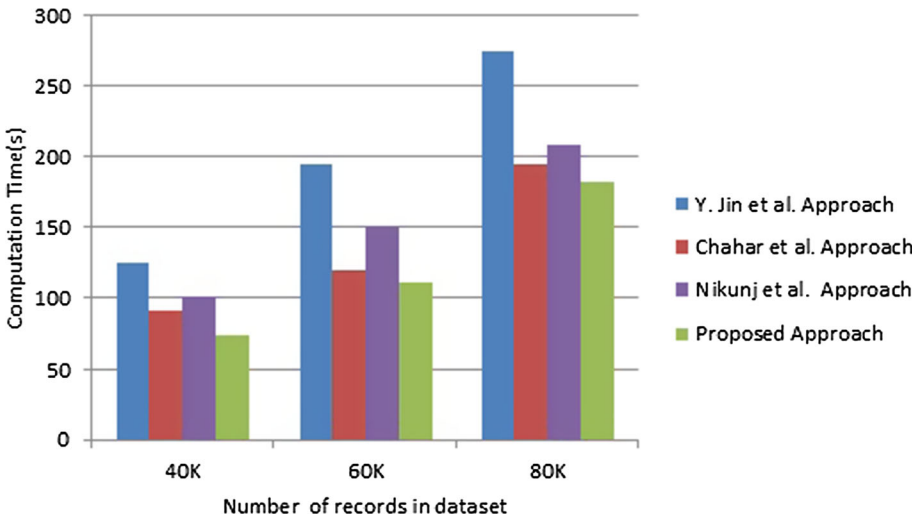

Fig. 9 Comparison of Computation Cost with Existing Approaches with Increasing Number of records for Heart Disease Dataset

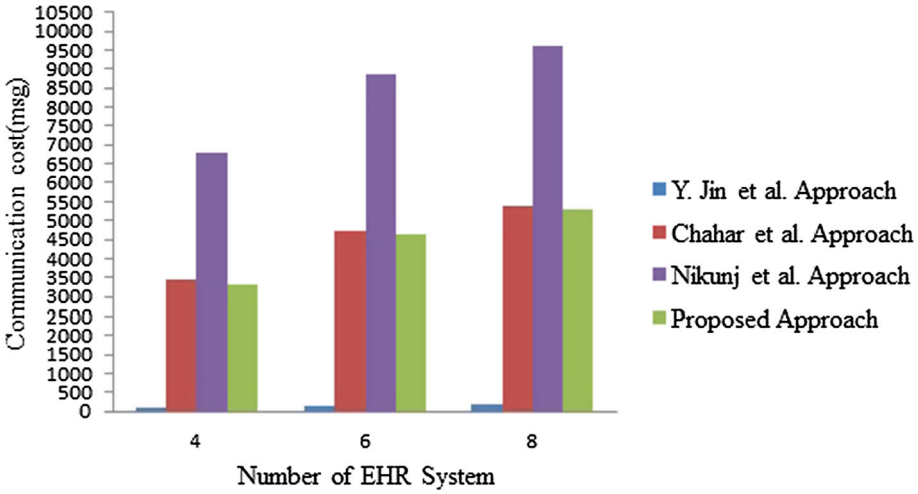

Fig. 10 Comparison of Communication Cost with Existing Approaches with Increasing Number of EHR System for Heart Disease Dataset

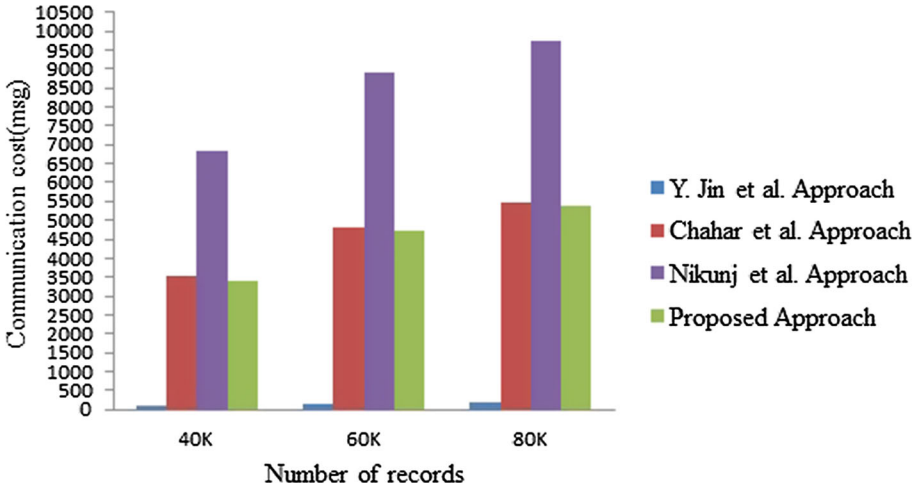

Fig. 11 Comparison of Communication Cost with Existing Approaches with Increasing Number of records for Heart Disease Dataset

a discussion of the privacy analysis of the proposed approach against any adversary:

Claim: *Proposed approach preserves the privacy under the hardness of solving CDH problem.*

Proof : After a complete execution of our approach central server obtains the global count of itemset as $Icount = \sum_{i=1}^{n} Icount_{(i)}$ for all itemsets. Some EHR systems or any malicious adversary with all message exchange among EHR systems collude and reveal the private value of some targeted EHR system. Here, it has been proved that such a coalition cannot reveal the private value of any EHR system. To reveal any EHR system's private data, an attacker has ciphertext as $C_i = E_y(g^{Icount_{(i)}+z_i}) = (g^{r_i}, g^{Icount_{(i)}+z_i} \cdot y^{r_i})$. For solving this ciphertext, an adversary has only known values are public parameters g, y. Value of z_i and $Icount_{(i)}$ is only known to EHR_i. If any group of EHR systems or adversary can solve the $g^{Icount_{(i)}+z_i}$ to reveal $Icount_{(i)}$, then they can resolve the CDH problem. The CDH problem is proved as hard to solve. Therefore, computing the private value, $Icount_{(i)}$ of any EHR_i by any malicious attacker is identical to the CDH problem. Hence, our approach preserves privacy under the hardness of solving the CDH problem.

Conclusion

By using collaborative association rule mining across distributed EHR systems, it was emphasized to improve disease prediction accuracy while protecting patient privacy. The proposed approach used Additive Homomorphic ElGamal Cryptosystem to preserve privacy while computing the global association rules from the different collaborative EHR systems. Experimental results show the benefits of the proposed approach using aggregated data from all EHR systems compared to individual EHR results. The possibility of combating the COVID-19 using the proposed approach is also discussed. In future, a detailed analysis will be done on COVID-19 patient data.

Funding This research received no specific funding from any funding agency.

Declarations

Conflict of interest Authors declares that there is no conflict of interest.

References

1. J. Nahar, T. Imam, K.S. Tickle, Y.-P.P. Chen, Association rule mining to detect factors which contribute to heart disease in males and females. Expert Syst. Appl. **40**(4), 1086–1093 (2013)
2. M. Heron, Deaths: leading causes for 2015. National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System **66**(5), 1–76 (2017)
3. F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clinic. **68**(6), 394–424 (2018)
4. "Covid-19 coronavirus pandemic," [Online] Available: https://www.worldometers.info/coronavirus/, [Accessed: 15-May-2020]
5. Z. Kakushadze, R. Raghubanshi, W. Yu, Estimating cost savings from early cancer diagnosis. Data **2**(3), 30 (2017)
6. Q. Alefan, A. Saadeh, R.J. Yaghan, Direct medical costs for stage-specific breast cancer: a retrospective analysis. Breast Cancer Manag. **9**(1), BMT33 (2020)
7. R. Simic, N. Ratkovic, V. D. Simic, Z. Savkovic, M. Jakovljevic, V. Peric, M. Pandrc, and N. Rancic, "Cost analysis of health examination screening program for ischemic heart disease in active-duty military personnel in the middle-income country. Front. Public Health. 9, 2021
8. V.J. Kadam, S.M. Jadhav, K. Vijayakumar, Breast cancer diagnosis using feature ensemble learning based on stacked sparse autoencoders and softmax regression. J. Med. Syst. **43**(8), 1–11 (2019)
9. H. Wang, B. Zheng, S.W. Yoon, H.S. Ko, A support vector machine-based ensemble algorithm for breast cancer diagnosis. Eur. J. Oper. Res. **267**(2), 687–699 (2018)
10. A.K. Dubey, U. Gupta, S. Jain, Analysis of k-means clustering approach on the breast cancer wisconsin dataset. Int. J. Computer Ass. Radiol. Surg. **11**(11), 2033–2047 (2016)
11. H. Asri, H. Mousannif, H. Al Moatassime, T. Noel, Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Sci. **83**, 1064–1069 (2016)
12. C.P. Utomo, A. Kardiana, R. Yuliwulandari, Breast cancer diagnosis using artificial neural networks with extreme learning techniques. Int. J. Adv. Res. Artif. Intell. **3**(7), 10–14 (2014)
13. G. I. Salama, M. Abdelhalim, and M. A.-e. Zeid, "Experimental comparison of classifiers for breast cancer diagnosis," in 2012 Seventh International Conference on Computer Engineering & Systems (ICCES). IEEE, 2012, pp. 180–185
14. G.I. Salama, M. Abdelhalim, M.A.-E. Zeid, Breast cancer diagnosis on three different datasets using multi-classifiers. Breast Cancer (WDBC) **32**(569), 2 (2012)
15. D. Lavanya, K.U. Rani, Ensemble decision tree classifier for breast cancer data. Int. J. Information Technol. Converg. Serv. **2**(1), 17 (2012)
16. A.M. Abdel-Zaher, A.M. Eldeib, Breast cancer classification using deep belief networks. Expert Syst. Appl. **46**, 139–144 (2016)
17. P.C. Tang, C.J. McDonald, Electronic health record systems. Biomed. Inform. **10**(4), 447–475 (2006)
18. M. Harahap, A. Husein, S. Aisyah, F. Lubis, and B. Wijaya, "Mining association rule based on the diseases population for recommendation of medicine need," in Journal of Physics: Conference Series, vol. 1007, no. 1. IOP Publishing, 2018, p. 012017
19. M. Tandan, Y. Acharya, S. Pokharel, M. Timilsina, Discovering symptom patterns of covid-19 patients using association rule mining. Computers Biol. Med. **131**, 104249 (2021)
20. W. Altaf, M. Shahbaz, A. Guergachi, Applications of association rule mining in health informatics: a survey. Artif. Intell. Rev. **47**(3), 313–340 (2017)
21. S.M. Kang, P.W. Wagacha, Extracting diagnosis patterns in electronic medical records using association rule mining. Int. J. Computer Appl. **108**(15), (2014)

22. S. Babu, E. Vivek, K. Famina, K. Fida, P. Aswathi, M. Shanid, M. Hena, "Heart disease diagnosis using data mining technique," in, international conference of electronics, communication and aerospace technology (ICECA), vol. 1. IEEE **2017**, 750–753 (2017)
23. A.M. Khedr, Z. Al Aghbari, A. Al Ali, M. Eljamil, An efficient association rule mining from distributed medical databases for predicting heart diseases. IEEE Access **9**, 15320–15333 (2021)
24. A. Yazdani, K.D. Varathan, Y.K. Chiam, A.W. Malik, W.A.W. Ahmad, A novel approach for heart disease prediction using strength scores with significant predictors. BMC Med. Inform. Decis. Mak. **21**(1), 1–16 (2021)
25. A.M. Shin, I.H. Lee, G.H. Lee, H.J. Park, H.S. Park, K.I. Yoon, J.J. Lee, Y.N. Kim, Diagnostic analysis of patients with essential hypertension using association rule mining. Healthcare inform. Res. **16**(2), 77–81 (2010)
26. S. Palaniappan, R. Awang, Intelligent heart disease prediction system using data mining techniques. Int. J. Healthcare Biomed. Res. **1**, 94–101 (2013)
27. C. Ordonez, Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions Information Technol. Biomed. **10**(2), 334–343 (2006)
28. N. Y. You, K. S. Ryu, J. H. Kim, and H. Y. J. Kang, "Association rule mining method to predict coronary artery disease: Knhanes," in Advances in Intelligent Information Hiding and Multimedia Signal Processing, vol. 211. Springer, 2021, p. 274
29. S.J. Lee, K.B. Cartmell, An association rule mining analysis of lifestyle behavioral risk factors in cancer survivors with high cardiovascular disease risk. J. Pers. Med. **11**(5), 366 (2021)
30. M. Karabatak, M.C. Ince, An expert system for detection of breast cancer based on association rules and neural network. Expert Syst. Appl. **36**(2), 3465–3469 (2009)
31. H.C. Koh, G. Tan et al., Data mining applications in healthcare. J. Healthcare Information Manag. **19**(2), 65 (2011)
32. C. Clifton, M. Kantarcioglu, J. Vaidya, Defining privacy for data mining. National Science Foundation Workshop on Next Generation Data Mining **1**(26), 199–204 (2002)
33. A. Gkoulalas-Divanis, G. Loukides, J. Sun, Publishing data from electronic health records while preserving privacy: a survey of algorithms. J. Biomed. Inform. **50**, 4–19 (2014)
34. A. Telikani, A.H. Gandomi, A. Shahbahrami, A survey of evolutionary computation for association rule mining. Information Sci. **524**, 318–352 (2020)
35. L. Zhang, W. Wang, Y. Zhang, Privacy preserving association rule mining: taxonomy, techniques, and metrics. IEEE Access **7**, 45032–45047 (2019)
36. D. Gunawan, Classification of privacy preserving data mining algorithms: a review. Jurnal Elektronika dan Telekomunikasi **20**(2), 36–46 (2020)
37. K. Nomura, Y. Shiraishi, M. Mohri, M. Morii, Secure association rule mining on vertically partitioned data using private-set intersection. IEEE Access **8**, 144458–144467 (2020)
38. "Breast cancer wisconsin (original) data set," [Online] Available: https://archive.ics.uci.edu/ml/ machine-learning-databases/ breast-cancer-wisconsin /breast-cancer-wisconsin.data, [Accessed: 28-May-2018]
39. "Heart disease dataset," [Online] Available: http: //archive.ics.uci.edu/ml/machine-learningdatabases/heart-disease/cleve.mod, [Accessed: 28-May-2018]
40. M. Kantarcioglu, C. Clifton, Privacy preserving distributed mining of association rules on horizontally partitioned data. IEEE Transactions Knowl. Data Eng. **9**, 1026–1037 (2004)
41. V. S. Verykios and A. Gkoulalas-Divanis, "A survey of association rule hiding methods for privacy," in Proceedings of Privacy Preserving Data Mining. Springer, 2008, pp 267–289
42. C.-W. Lin, T.-P. Hong, H.-C. Hsu, Reducing side effects of hiding sensitive itemsets in privacy preserving data mining. Scientific World J. **2014**, 267–289 (2014)
43. M. B. Malik, M. A. Ghazi, and R. Ali, "Privacy preserving data mining techniques: current scenario and future prospects," in Proceedings of Third International Conference on Computer and Communication Technology (ICCCT). IEEE, 2012, pp. 26–32
44. N.R. Nanavati, P. Lalwani, D.C. Jinwala, Analysis and evaluation of schemes for secure sum in collaborative frequent itemset mining across horizontally partitioned data. J. Eng. **2014**, 110–120 (2014)
45. N. Domadiya, U.P. Rao, Privacy-preserving association rule mining for horizontally partitioned healthcare data: a case study on the heart diseases. Sādhanā **43**(8), 127 (2018)
46. X. C. Nguyen, H. B. Le, and T. A. Cao, "An enhanced scheme for privacy-preserving association rules mining on horizontally distributed databases," in Proceedings of IEEE International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF). IEEE, 2014, pp. 1–4
47. S. Mewada, Data mining-based privacy preservation technique for medical dataset over horizontal partitioned. Int. J. E-Health Med. Commun. (IJEHMC) **12**(5), 50–66 (2021)
48. N. Domadiya, U.P. Rao, Privacy preserving association rule mining on distributed healthcare data: Covid-19 and breast cancer case study. SN Computer Sci. **2**(6), 1–9 (2021)
49. J. S. Vaidya, "Privacy preserving data mining over vertically partitioned data," Ph.D. dissertation, West Lafayette, IN, USA, 2004
50. N.R. Nanavati, D.C. Jinwala, A novel privacy-preserving scheme for collaborative frequent itemset mining across vertically partitioned data. Secur. Commun. Netw. **8**(18), 4407–4420 (2015)
51. Z. Xu and X. Yi, "Classification of privacy-preserving distributed data mining protocols," in Proceedings of Sixth International Conference on Digital Information Management. IEEE, 2011, pp. 337–342
52. M. Yogasini, B. Prathibha, Secure association rule mining on vertically partitioned data using fully homomorphic encryption. ICTACT J. Soft Comput. **11**(4), 2424–2428 (2021)
53. N. Domadiya, U.P. Rao, Privacy preserving distributed association rule mining approach on vertically partitioned healthcare data. Procedia Computer Sci. **148**, 303–312 (2019)
54. R. Agrawal, R. Srikant et al., "Fast algorithms for mining association rules," in Proceeding of 20th international conference on very large data bases, VLDB, vol. 1215, 1994, pp. 487–499
55. T. ElGamal, A public key cryptosystem and a signature scheme based on discrete logarithms. IEEE Transactions Information Theory **31**(4), 469–472 (1985)
56. R. Cramer, R. Gennaro, B. Schoenmakers, A secure and optimally efficient multi-authority election scheme. Eur. Transactions Telecommun. **8**(5), 481–490 (1997)
57. D. Burdick, M. Calimlim, and J. Gehrke, "Mafia: A maximal frequent itemset algorithm for transactional databases," in Proceedings. 17th International Conference on Data Engineering. IEEE, 2001, pp. 443–452
58. H. Chahar, B.N. Keshavamurthy, C. Modi, Privacy-preserving distributed mining of association rules using elliptic-curve cryptosystem and shamir's secret sharing scheme. Sādhanā. **42**(12), 1997–2007 (2017)
59. Y. Jin, C. Su, N. Ruan, and W. Jia, "Privacy-preserving mining of association rules for horizontally distributed databases based on fp-tree," in International Conference on Information Security Practice and Experience. Springer, 2016, pp. 300–314

60. "Cleveland heart disease data details," [Online] Available: http://archive.ics.uci.edu/ml/ machine-learning-databases/heart-disease/heart-disease.names, [Accessed: 28-May-2016]
61. P. Melin, J.C. Monica, D. Sanchez, O. Castillo, Analysis of spatial spread relationships of coronavirus (covid-19) pandemic in the world using self organizing maps. Chaos Solitons Fractals **138**, 109917 (2020)
62. Q.-X. Ma, H. Shan, H.-L. Zhang, G.-M. Li, R.-M. Yang, J.-M. Chen, Potential utilities of mask-wearing and instant hand hygiene for fighting sars-cov-2. J. Med. Virol. **92**(9), 1567–1571 (2020)
63. "Q/a on coronaviruses (covid-19)," [Online] Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses:text=symptoms, [Accessed: 20-April-2020]
64. T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, E. B. Hsu, S. Yang, and P. Eklund, "Artificial intelligence in the battle against coronavirus (covid-19): a survey and future research directions," arXiv preprint arXiv:2008.07343, 2020
65. A. Naz, F. Shahid, T.T. Butt, F.M. Awan, A. Ali, A. Malik, Designing multi-epitope vaccines to combat emerging coronavirus disease 2019 (covid-19) by employing immuno-informatics approach. Front. Immunol. **11**, 1663 (2020)
66. A. Kumar, K. Sharma, H. Singh, S.G. Naugriya, S.S. Gill, R. Buyya, A drone-based networked system and methods for combating coronavirus disease (covid-19) pandemic. Future Gener. Computer Syst. **115**, 1–19 (2021)
67. L. Muhammad, M.M. Islam, S.S. Usman, S.I. Ayon, Predictive data mining models for novel coronavirus (covid-19) infected patients' recovery. SN Computer Sci. **1**(4), 1–7 (2020)
68. A.K. Arshadi, J. Webb, M. Salem, E. Cruz, S. Calad-Thomson, N. Ghadirian, J. Collins, E. Diez-Cecilia, B. Kelly, H. Goodarzi et al., Artificial intelligence for covid-19 drug discovery and vaccine development. Front. Artific. Intell. **3**, 65 (2020)