SN Computer Science (2021) 2:418
https://doi.org/10.1007/s42979-021-00801-7

ORIGINAL RESEARCH

Privacy Preserving Association Rule Mining on Distributed Healthcare Data: COVID-19 and Breast Cancer Case Study

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

Received: 25 May 2020 / Accepted: 29 July 2021 / Published online: 18 August 2021

Abstract

Association rule mining can be used in healthcare data mining to provide solutions to life-threatening diseases like recent COVID-19. Due to healthcare data privacy concerns, privacy preserving distributed healthcare data mining becomes the primary focus of medical science research. Recently, Chahar et al. (Sādhanā 42:1997–2007, 2017) proposed privacy preserving distributed association rule mining scheme with insecure communication channels. They used the concept of an elliptic curve-based paillier cryptosystem to achieve privacy, authenticity, and integrity. We observed some security vulnerabilities in their privacy preserving association rule mining scheme when implemented with insecure communication channels. We observed that the security vulnerabilities will result in the disclosure of private data of sites (or participants). Furthermore, we propose a secure version of their scheme to solve the security vulnerabilities with insecure communication channels. Theoretical and experimental analysis shows that the proposed scheme has almost equal computation and communication complexities with better securities. A case study on the effectiveness of the proposed approach in combating COVID-19 coronavirus and Breast Cancer is also discussed.

Keywords Healthcare data mining · Privacy · Privacy preserving data mining · Brest cancer · COVID-19

Introduction

Recently, COVID-19 or coronavirus and cancer disease among various life-threatening diseases have received a significant attention of medical researchers. Across the globe, life-threatening diseases are the main focus of study in medical research due to its impact on the life of humans [1]. US Department of Health and Human Services reported 45.3% of total death because of cancer and heart disease, as shown in Fig. 1. American Cancer Society listed breast cancer as a leading type of cancer among various types of cancer, claiming millions of women lives in the United States. Figure 1 shows the expected number of different types of cancer cases in the United States in 2018. Breast cancer is having the highest number of estimation of 19% among different types of cancer. The novel disease started since 2019 known as Coronavirus disease (COVID-19) reported total of 4,527,815 cases around the world with death of 303,438 till 15th May-2020 [2]. Considering this life-threatening diseases with huge mortality rate, early detection of disease by analyzing patient's symptoms is essential to save more lives.

Proper treatment and recovery from these life-threatening disease prerequisite the disease identification at an early stage. The diagnostic system of cancer and heart disease are costly, prone to error and time-consuming. In the past, prediction of disease is often based on the expertise of physician rather than symptoms patterns hidden in healthcare data. Hence, this may cause the error in the diagnosis of disease resulting in unnecessary medical treatment which increases the healthcare expenses and affects the quality of healthcare services to patients. The trends of electronic healthcare record (EHR) system in major hospitals are increasing. It stores the enormous data about patient's information [4]. Data collected at hospitals can be used for healthcare research and improving the healthcare services using data mining. Association rule mining is one of the famous data mining technique for finding co-relation of disease and symptoms. Various applications of association rule mining in healthcare domain are as follows: (1)

✉ Nikunj Domadiya
domadiyanikunj002@gmail.com

Udai Pratap Rao
upr@coed.svnit.ac.in

[1] Department of Computer Engineering, Sardar Vallabhbhai National Institute of Technology, Surat 395007, India

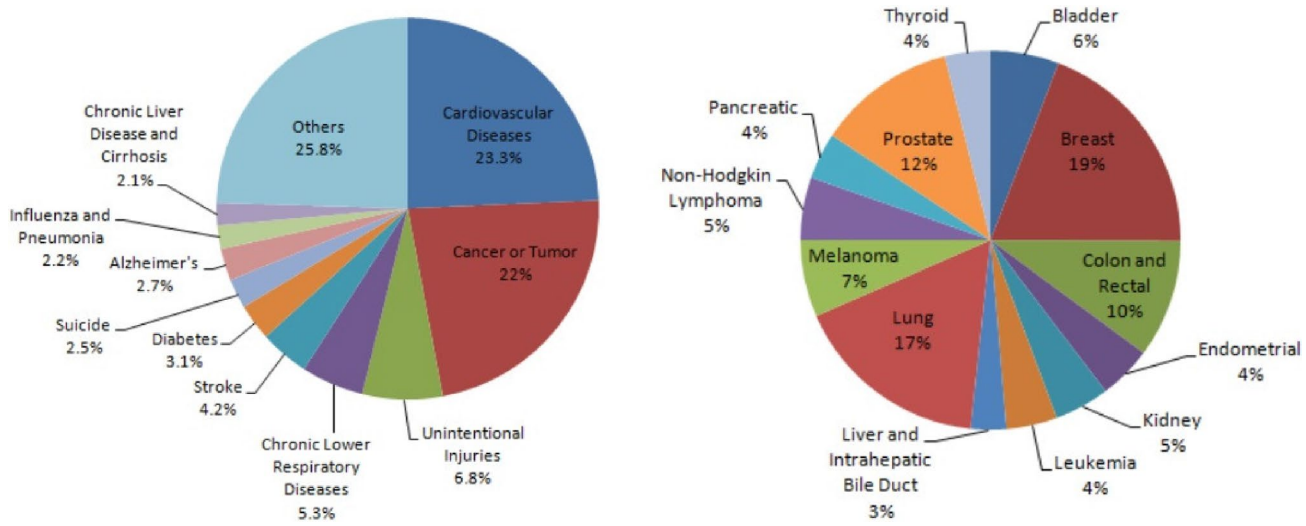

Fig. 1 United States Health Statistics [3]

predicting disease based on symptoms of a patient, (2) identifying the most effective treatment for diseases, (3) medical prescription, discovering the reaction of drug, (4) medical fraud identification. These domains can be improved using data mining [5–7]. Association rule mining technique on healthcare data results in form of *IF–THEN* rules (For Ex. IF *Bland chromatin=3* and *Bare nuclei=1* and *Mitoses=1* then *class=benign*). Physicians or medical experts can assuredly interpret these rules. Hence, this technique is very renowned among medical researchers and physicians for identifying the disease or preferred treatment of specific disease based on the patient's symptoms. That can help in reducing healthcare expenses and suggest the standard treatment for particular diseases. Therefore, Healthcare treatment and diagnosis becomes faster and efficient [8].

Currently, association rule mining on healthcare data is limited to a single hospital with its own EHR system [9]. Single EHR system has limited patient's data. Hence, accuracy regarding confidence of association rule mining on a single EHR system's data is lower. Finding the relationship between harmful diseases(COVID-19, cancer, heart disease, etc.) and the patient's symptoms requires more precise association rules [7]. The accuracy/confidence of association rule mining can be improved by aggregating the data from all EHR systems to some central data mining server. Association rule mining on central server's aggregate data gives better accuracy compared to results of single EHR system's data. Privacy of data collected at local EHR system must be preserved because of social, economic, and psychological issues may occur to patients when their healthcare data are revealed [10, 11]. Collaboration of different EHR systems for accurate data mining requires the sharing of local EHR system's data while maintaining the privacy. Therefore, medical research focused on privacy preserving association rule mining on distributed healthcare data. Distributed data are either horizontally partitioned or vertically partitioned. The schema of EHR system at major hospitals is the same as they follow the standard rules related to patient's information that must be stored in hospitals. Hence, we have considered horizontally partitioned data among the collaborative EHR systems. Based on these reviews, this research focuses on generating accurate global association rules for diagnosis of specific disease (e.g., COVID-19 and cancer disease) from the distributed healthcare data while maintaining the privacy of EHR systems.

The increasing trends of extracting interesting patterns by collaborative data mining requires the sharing of data among collaborative participants. Collaborative data mining in health, medicine, government agencies, and business-related applications raises the privacy issues on disclosing the private data. Association rule mining is very eminent data mining technique in various applications. Therefore, privacy preserving distributed association rule mining (PPDARM) became a key research direction.

The PPDARM approach proposed by Clifton and Kantarcioglu [12] used a random number addition which works without public key homomorphism. However, this scheme is not secure against the collusion of involving parties and discloses the private data to an external attacker if it captures the communication among parties. This scheme works better if all parties use the secure communication channel, but it increases the computation cost. In [12–17], authors use the homomorphic encryption technique [18] for a secure sum [19]. It has higher computation cost because of complex public key operation. In [20], approaches use the trusted third party model for a secure sum. However, the privacy of

individual totally depends on the behavior of TTP; if TTP compromises, then individual privacy is not maintained. In [21–23], author proposed a scheme based on Shamir secret sharing. However, it has higher communication cost as it counts the frequency for each candidate itemset. Chahar et al. [24] proposed an approach homomorphic property of elliptic curve-based paillier cryptosystem [25] for PPDARM in horizontally partitioned data with insecure communication channel. However, after analyzing the scheme of Chahar et al. [24], we identified some security flaws in protecting the privacy of participants. The private data of each participants can be easily learned by miner participant.

Next, "Review of H. Chahar et al. Scheme" discuss the Chahar et al. [24] scheme. "Security Flaws in H. Chahar et al.'s Protocol [24]" presents the security flaws in Chahar et al.'s [24] scheme. "Proposed Secure version of H. Chahar et al.'s Protocol [24]" presents the proposed solution for preserving the privacy. Analysis of propose scheme is discussed in "Analysis of Proposed Scheme". Case study on effectiveness of propose approach for Breast Cancer and COVID-19 is discussed in "Case Study: Effectiveness of Proposed Scheme for Breast Cancer and COVID-19".

Review of H. Chahar et al.'s Scheme

In this section, we review an efficient privacy preserving distributed mining of association rules scheme, recently proposed by Chahar et al. [24].

Chahar et al. [24] proposed two protocols for securely mining global association rules over horizontally partitioned data. The first protocol is based on elliptic curve based paillier cryptosystem [25] and the second protocol is based on shamir's secret sharing [26]. The first protocol uses the homomorphic property of elliptic curve-based paillier cryptosystem for securely computing the sum of itemset count from all distributed sites. All messages are exchanged after encryption and signing to achieve privacy, authenticity, and integrity. The second protocol tries to prevent the collusion among the sites using shamir's secret sharing.

Privacy preserving distributed association rule mining scheme can be considered as follows: n sites $\{\text{site}_1, \text{site}_2, \text{site}_3, \ldots, \text{site}_n\}$ with distributed database DB in such a way that site_i stores the database DB_i where $\text{DB} = \sum_{i=1}^{n} \text{DB}_i$. Database DB is considered as horizontally partitioned among n sites with all sites having the same schema but different number of transactions. One site is considered as combiner, second site as miner and remaining all sites are collaborative sites, as shown in Fig. 2. Certificate authority is responsible for generating an elliptic curve-based paillier secret and public keys for all collaborative sites. Homomorphic property of an elliptic curve-based paillier cryptosystem helps to compute global count of itemset securely.

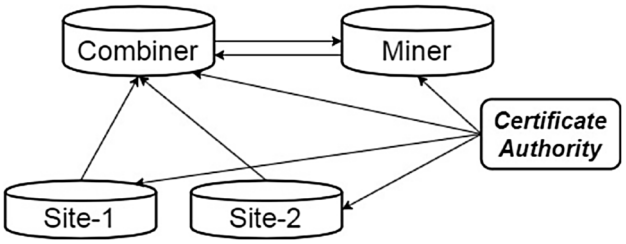

Fig. 2 Communication model of H. Chahar et al.'s protocol

The scheme proposed by Chahar et al. [24] works with the semi-honest model. In semi-honest model, collaborative site does not collude with other sites, and follows the proposed protocol/algorithm and can try to infer the private value of other sites from the protocol data exchanges. Their protocol works with insecure communication channel among sites. Hence, any adversary can snoop all communication among sites.

Lemma 1 *Elliptic curve-based paillier homomorphic property* [25] *is used to compute global support count of an itemset P with n sites having local support count* $m_1, m_2, m_3, \ldots, m_n$ *at respective site as follows*: *For any n values* $m_1, m_2, m_3, \ldots, m_n$,

Encryption: $E(m_1) * E(m_2) * E(m_3) * \cdots * E(m_n) = E(m_1 + m_2 + m_3 + \cdots + m_n)$

Decryption: $D(E(m_1) * E(m_2) * E(m_3) * \cdots * E(m_n)) = m_1 + m_2 + m_3 + \cdots + m_n$.

Lemma 2 $\bigcup_{i=1}^{n} \text{MFI}_i$ *determines all global frequent itemsets* (MFI_i *is local MFI at* site_i) [27].

The protocol proposed in [24] works in three phases. Next, we discuss each phase in brief.

Phase 1: Computation of global maximal frequent itemset (MFI) from all sites

- Public and secret keys for each collaborative site are generated by certificate authority using elliptic-curve paillier public key cryptosystem [25]. The key pair generated for sites $\{\text{site}_1, \text{site}_2, \text{site}_3(\text{combiner}), \text{site}_4(\text{miner})\}$ are $\{(\text{pk}_1, \text{sk}_1), (\text{pk}_2, \text{sk}_2), (\text{pk}_3, \text{sk}_3), (\text{pk}_4, \text{sk}_4)\}$, respectively. Hence, each site has public keys of all sites and its own secret key.
- Each site computes its own maximal frequent itemsets (MFI), encrypts it using the miner's public key, and signs the encrypted data with its own secret key. Each site (except miner and combiner) then sends this signed message to combiner.

- Combiner authenticates all the received messages using the public key of respective sites. It combines its own encrypted MFI with all received messages and shuffle it. Next, it signs the combined message using its own secret key and sends it to miner.
- Miner authenticates the received message and decrypt it to find the local maximal frequent itemset (MFI) from all the sites. It combines its own MFI with received MFI and finds the global MFI. Global MFI is finally shared with all the sites by miner.

Phase 2: Computation of global support count of candidate itemset from all sites

- The support counts of candidate itemsets are computed by each site which are obtained from all the subset of global MFI. For each candidate itemset P, each site (except miner) encrypts the local support count of P using miner's public key (pk_4), and sends it to the combiner with its own signature on encrypted message. Encrypted message of candidate itemset P at any $site_i$ is symbolized by $E(P.sup_i)$.
- Combiner authenticates the received message using the public key of respective sites and computes $E(P.sup_1) * E(P.sup_2) * E(P.sup_{combiner})$. Combiner sends this message to miner with its own signature.
- Miner authenticates the received message using combiner's public key. Next, it decrypts the message using its own secret key as $D(E(P.sup_1) * E(P.sup_2) * E(P.sup_{combiner})) = (P.sup_1 + P.sup_2 + P.sup_{combiner})$. It computes the global support count of itemset P by adding its own support count as: $P_{globalsup} = (P.sup_1 + P.sup_2 + P.sup_{combiner}) + P.sup_{miner}$

Phase 3: Computation of global database size and association rules

- Size of global database DB is computed as $|DB| = \sum_{i=1}^{n} |DB_i|$, where $|DB_i|$ indicates the database size at $site_i$.
- Finally, miner computes global association rules from all the global frequent itemsets and global database size. Miner shares global association rules with all the collaborative sites.

Security Flaws in H. Chahar et al.'s Protocol [24]

Chahar et al. [24] declared that their scheme achieves privacy in the semi-honest model with insecure communication channels. However, we observed that the private data of all the sites can be disclosed in Chahar et al.'s protocol [24] by miner. Next, we discuss the security flaws in [24] by the miner.

In the second phase of Chahar et al.'s protocol [24], each $site_i$ $(1 \le i \le n-2)$ sends the encrypted support count of an itemset P as $E(P.sup_i)$ to combiner with signature of respective $site_i$. Due to insecure communication channel, miner could also eavesdrop the $E(P.sup_i)$ from all sites. Miner decrypts $E(P.sup_i)$ and learns the private data $P.sup_i$, $(1 \le i \le n-2)$ of $site_i$. Hence, privacy of all sites gets jeopardized. Miner could also learn the private data of combiner by subtracting the $\prod_{i=1}^{n-2} P.sup_i + P.sup_{miner}$ from $P_{globalsup}$. Hence, miner could jeopardize the private information of all sites. As a result, Chahar et al.'s protocol [24] fails to preserve privacy of participants in insecure communication environment.

In this paper, we propose the solution to preserve the privacy of participants in Chahar et al.'s protocol [24] with insecure communication environment.

Proposed Secure Version of H. Chahar et al.'s Protocol [24]

We have modified phase-1 and phase-2 of Chahar et al.'s protocol [24]. Proposed scheme works with one site as Miner and second site as Combiner and the remaining as normal sites for collaboration, as shown in Fig. 2. Symbols used in the proposed approach are defined in Table 1.

Computation of an itemset *P's* global count in proposed scheme works in three phases as follows:

Phase 1: Computation of global maximal frequent itemset (MFI) from all sites

- Public and secret keys for all collaborative sites are generated by certificate authority using elliptic-curve paillier public key cryptosystem. The key pair generated for sites $\{site_{combiner}, site_{miner}, site_1, site_2, \ldots, site_{n-2}\}$ are $\{(pk_c, sk_c), (pk_m, sk_m), (pk_1, sk_1), (pk_2, sk_2), \ldots, (pk_{n-2}, sk_{n-2})\}$, respectively. Hence, each site has public keys of all sites and its own secret key.
- Certificate authority sends Z to all sites including combiner (except miner), where Z is an integer greater than total number of transactions of all sites.

Table 1 Notations

Notation	Descriptions
$site_i$	Collaborative participant/site i
pk_i, sk_i	Public and private key pair of participant i
Z	Integer number shared by Certificate Authority to sites
$P.sup_i$	He support count of candidate itemset P at $site_i$
$\|DB_i\|$	Dataset size in number of records at $site_i$

- Each site computes its own maximal frequent itemsets (MFI), encrypts it using the miner's public key, and signs the encrypted data with its own secret key. Each site (except miner and combiner) sends this signed message to combiner.
- Combiner authenticates all the received messages using the public keys of respective sites. It combines its own encrypted MFI with all received messages and shuffles it. Next, it signs the combined message using its own secret key and sends it to miner.
- Miner authenticates the received message and decrypts it to find the local maximal frequent itemset (MFI) from all sites. It combines its own MFI with received MFI and finds the global MFI. Global MFI is shared with all sites by miner.

Phase 2: Computation of global support count of candidate itemset from all sites

- Each site_i (except miner) computes $(P.\text{sup}_i + R_i * Z)$, where R_i is the random number selected by site_i and $P.\text{sup}_i$ is the support count of candidate itemset P at site_i. Then, site_i encrypts it using miner's public key (pk_m) as $E(P.\text{sup}_i + R_i * Z)$ and sends it to combiner with its own signature on encrypted message.
- Combiner authenticates the received message using the public key of respective sites and computes $E(P.\text{sup}_{\text{Combiner}}) = E(P.\text{sup}_{\text{Combiner}} + R_{\text{Combiner}} * Z) * \prod_{i=1}^{n-2} E(P.\text{sup}_i + R_i * Z)$. Combiner sends this message to miner with its own signature.
- Miner authenticates the received message using combiner's public key. Then, it decrypts the message using its own secret key as $D(E(P.\text{sup}_{\text{Combiner}})) = (P.\text{sup}_{\text{Combiner}} + R_{\text{Combiner}} * Z) + (P.\text{sup}_1 + R_1 * Z) + (P.\text{sup}_2 + R_2 * Z) + \cdots + (P.\text{sup}_{n-2} + R_{n-2} * Z) = (P.\text{sup}_{\text{Combiner}} + P.\text{sup}_1 + P.\text{sup}_2 + \cdots + P.\text{sup}_{n-2}) + (R_{\text{Combiner}} + R_1 + R_2 + \cdots + R_{n-2}) * Z$
- Miner combines its own local support count of candidate itemset P with decrypted value as $((P.\text{sup}_{\text{Miner}} + P.\text{sup}_{\text{Combiner}} + P.\text{sup}_1 + P.\text{sup}_2 + \cdots + P.\text{sup}_{n-2}) + (R_{\text{Combiner}} + R_1 + R_2 + \cdots + R_{n-2}) * Z)$ and sends it to combiner with its own signature.
- Combiner authenticates the received message and computes the global support count $P_{\text{globalsup}}$ for candidate itemset P as follows:

$$
\begin{aligned}
P_{\text{globalsup}} &= ((P.\text{sup}_{\text{Miner}} + P.\text{sup}_{\text{Combiner}} + P.\text{sup}_1 + P.\text{sup}_2 + \cdots + P.\text{sup}_{n-2}) \\
&\quad + (R_{\text{Combiner}} + R_1 + R_2 + \cdots + R_{n-2}) * Z) \mod Z \\
&= (P.\text{sup}_{\text{Miner}} + P.\text{sup}_{\text{Combiner}} + P.\text{sup}_1 + P.\text{sup}_2 + \cdots + P.\text{sup}_{n-2})
\end{aligned}
$$

Combiner sends $P_{\text{globalsup}}$ to all sites.

Phase 3: Computation of global database size and association rules

- Size of global database DB is computed as $|\text{DB}| = \sum_{i=1}^{n} |\text{DB}_i|$, where $|\text{DB}_i|$ indicates the database size at site_i.
- Finally, all sites compute global association rules from all global frequent itemset and global database size.

The proposed approach securely computes global association rules and ensures that the privacy of all sites with insecure communication environment is preserved. All messages among the sites are exchanged after encryption and signature by the respective sites to achieve integrity and authenticity.

Analysis of Proposed Scheme

Security Analysis of proposed Approach

Privacy of Site Participants: Each site_i sends the private data in encrypted form as $E(P.\text{sup}_i + R_k * Z)$ to combiner. Combiner is impotent to decrypt and find $P.\text{sup}_i$ for any site_i. Miner can eavesdrop $E(X.\text{sup}_k + R_k * Z)$ from insecure communication channel and can compute the $(P.\text{sup}_i + R_i * Z)$ using its secret key. Miner is unable to infer the private value $P.\text{sup}_i$ from $(P.\text{sup}_i + R_i * Z)$ as R_i and Z is not known to miner. Hence, privacy of all sites is preserved.

Privacy of Combiner: Miner receives $((P.\text{sup}_{\text{Combiner}} + P.\text{sup}_1 + P.\text{sup}_2 + \cdots + P.\text{sup}_{n-1}) + (R_{\text{Combiner}} + R_1 + R_2 + \cdots + R_{n-2}) * Z)$ as a single value. Hence, miner cannot infer the private value of combiner from received value and privacy of combiner is preserved.

Privacy of Miner: Miner sends $((P.\text{sup}_{\text{Miner}} + P.\text{sup}_{\text{Combiner}} + P.\text{sup}_1 + P.\text{sup}_2 + \cdots + P.\text{sup}_{n-2}) + (R_{\text{Combiner}} + R_1 + R_2 + \cdots + R_{n-2}) * Z)$ as a single value to combiner. Combiner cannot infer any private data of miner or other sites from the received value and privacy of miner is preserved.

Privacy Against External Adversary: All messages' exchanges among collaborative sites are encrypted and signed by respective sites. Hence, privacy of all sites is preserved against an external adversary.

Correctness Analysis

In semi-honest model, all sites follow the protocol. Combiner receives $E(P.\sup_i + R_i * Z)$ from site$_i$, where $(1 \le i \le n-2)$. Combiner computes $E(P.\sup_{\text{Combiner}}) = E(P.\sup_{\text{Combiner}} + R_{\text{Combiner}} * Z) * \prod_{i=1}^{n-2} E(P.\sup_i + R_i * Z)$ and sends it to miner. Miner decrypts the received value, adds its own local support count as $((P.\sup_{\text{Miner}} + P.\sup_{\text{Combiner}} + P.\sup_1 + P.\sup_2 + \cdots + P.\sup_{n-2}) + (R_{\text{Combiner}} + R_1 + R_2 + \cdots + R_{n-2}) * Z)$, and sends it to combiner. Combiner computes global support count as $P_{\text{globalsup}} = ((P.\sup_{\text{Miner}} + P.\sup_{\text{Combiner}} + P.\sup_1 + P.\sup_2 + \cdots + P.\sup_{n-2}) + (R_{\text{Combiner}} + R_1 + R_2 + \cdots + R_{n-2}) * Z) \mod Z$. It computes the correct global support count from all sites using the following Eq. (1):

$$\begin{aligned}&((P.\sup_{\text{Miner}} + P.\sup_{\text{Combiner}} + P.\sup_1 + P.\sup_2 + \cdots + P.\sup_{n-2}) \\ &\quad + (R_{\text{Combiner}} + R_1 + R_2 + \cdots + R_{n-2}) * Z) \mod Z \\ &= (P.\sup_{\text{Miner}} + P.\sup_{\text{Combiner}} + P.\sup_1 + P.\sup_2 + \cdots + P.\sup_{n-1}).\end{aligned} \tag{1}$$

Here, $(P.\sup_{\text{Miner}} + P.\sup_{\text{Combiner}} + P.\sup_1 + P.\sup_2 + \cdots + P.\sup_{n-1}) < Z$. Therefore, the proposed scheme always computes the correct global candidate itemset support count.

Theoretical Analysis

Table 2 shows the comparative analysis of proposed and some existing approaches in terms of external adversary attack, privacy and communication cost. In distributed association rule mining with n collaborative sites, $(n-1)$ sites (Combiner and $n-2$ sites) send their local maximal frequent itemset (MFI) to miner incur communication cost of $O(n)$. Miner computes global MFI and shares it with all other $(n-1)$ sites that incurs communication cost of $O(n)$. Global MFI with $(2^m - 1)$ subsets or candidate itemsets, with each $(n-1)$ site sharing local support count of candidate itemset, incurs communication cost of $(2^m - 1)(n-1) = O(n * 2^m)$. The computation overhead in proposed approach compared to H. Chahar et al.'s protocol [24] is once extra multiplication of $R_i * Z$ with $P.\sup_i$ at *phase-2* and one encryption at miner site in *phase-2* of proposed approach. The communication overhead is one message exchange from miner to combiner in *phase-2* of proposed approach. The contribution of this overhead is negligible in the overall execution time. In phase 3, $(n-1)$ sites send the local database size to miner incur the cost of $O(n)$. As a result, total communication cost is $(2^m - 1)(n-1) + 3(n-1) = O(2^m * n)$.

Table 2 Comparative analysis of proposed approach with existing schemes

References	Cryptography technique	Approach	Prevention of external adversary attack	Preserve privacy with insecure communication channel	Communication cost
Hussein et al. [28]	RSA Cryptography	Use of Apriori-Tid algorithm for association rule mining and efficient communication model for lower communication cost	No	No	$O(n * 2^m)$
Nguyen et al. [29]	Paillier homomorphic encryption	Maximal frequent itemset (MFI) technique for lower communication cost and paillier homomorphic encryption for security with lower computation cost	Yes	No	$O(n * 2^m)$
Chahar et al. [24]	Elliptic curve-based paillier homomophic cryptosystem	Computing global support count using elliptic curve-based paillier homomorphic cryptosystem for security and all message exchanges with digital signature for authenticity	Yes	No	$O(n * 2^m)$
Proposed scheme	Elliptic curve-based paillier homomophic cryptosystem	Computing global support count using elliptic curve-based paillier homomorphic cryptosystem with insecure communication channels and all message exchanges with digital signature for authenticity	Yes	Yes	$O(n * 2^m)$

Case Study: Effectiveness of Proposed Scheme for Breast Cancer and COVID-19

First, we elucidate symptoms related to breast cancer along with detail description of the dataset used in this analysis. We implemented our approach using NetBeans with system configuration of Intel Core i5 2.1 GHz CPU and 4 GB RAM. We have duplicated the total records to increase the total number of records of 80K. All record are randomly divided among the all collaborative EHR system.

Analyzing the Breast Cancer Disease

One of the most dangerous types of cancer disease is breast cancer disease. To analyze the symptoms correlated to breast cancer and improving the prediction accuracy based on symptoms, Wisconsin breast cancer dataset is used which is publicly available on the UCI repository [30].

Wisconsin Breast Cancer Dataset Details

In Wisconsin breast cancer dataset, a total of 32 different attributes are available. Out of 32, 10 significant correlated attributes are considered in this analysis. Two class labels *benign* and *malignant* indicate the breast cancer state. All attributes are in the range 1–10 assigned by the pathologist while testing the FNA (Fine Needle Aspirate) tissue sample in breast cancer. The attributes of Wisconsin breast cancer dataset with details are illustrated in Table 3.

Experiment Results Analysis

In experimental result analysis, association rules with RHS value *class= benign or malignant* are considered for accuracy analysis. Accuracy (confidence) of association rules using proposed approach and traditional local data mining at

Table 3 Attribute details 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

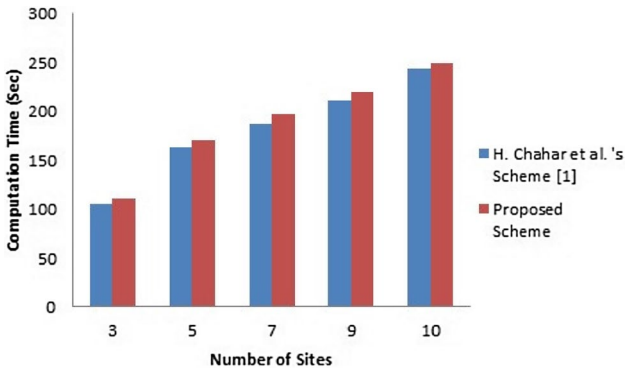

Fig. 3 Performance comparison of Chahar et al.'s protocol [24] and proposed scheme

Table 4 Prediction confidence/accuracy (%) at each EHR system and central data mining server for breast cancer

Association rules	Accuracy/confidence (%)				
	EHR1 (%)	EHR2 (%)	EHR3 (%)	EHR4 (%)	Central data mining server (proposed approach) (%)
Bland chromatin=3 and Bare nuclei=1 and Mitoses=1→ class=bengin	89	93	95	92	98
Uniformity of cell shape=1 and Uniformity of Cell Size=1→ class=benign	92	91	88	92	99
Bare nuclei=7 and Single epithelial cell size=1 and Normal nucleoli=2→class=benign	91	92	89	93	98
Uniformity of cell shape=10 and uniformity of Cell Size=8 Marginal adhesion=10→class=malignant	93	94	90	92	99
Bare nuclei=10 and uniformity of Cell Size=10→class=malignant	90	89	94	88	97
Bare nuclei=10 and Mitoses=2 and Normal nuclei=1→ class=malignant	87	85	90	91	96

each EHR system is illustrated in Table 4. The experimental result shows that the proposed approach has improved the accuracy of predicting breast cancer disease compared to any single EHR system.

The confidence of association rule indicates the prediction accuracy of breast cancer disease. As shown in Table 4, breast cancer prediction at each EHR system has lower precision compared to central data mining server results. Hence, the proposed scheme benefits every participant EHR system and physicians or medical researchers by allowing them access to global results which helps them to improve the healthcare services.

The proposed scheme and Chahar et al.'s protocol [24] are executed with three sites to ten sites. Comparison of both the schemes is shown in Fig. 3 with different number of sites. As shown in Fig. 3, computation cost of proposed scheme is almost equal compared to Chahar et al.'s protocol [24]. We have also added one extra step in the second phase of proposed secure scheme at combiner. It increases negligible computation and communication cost compared to Chahar et al.'s protocol [24]. Therefore, the proposed scheme is more secure with almost equal computation and communication cost.

Combating the Coronavirus Pandemic Using Proposed Approach

Coronaviruses are a large family of viruses that cause a range of illnesses from the common cold to Severe Acute Respiratory Syndrome (SARS-CoV). The new strain affecting tens of thousands across the globe is known as Novel Coronavirus (nCoV) and it causes the Coronavirus Disease also known as COVID-19.

This new disease and the propagating virus were relatively unknown before the outbreak happened from its origin at Wuhan province, China in December 2019. Hence, it was given the name COVID-19 due to the year it originated.

The most common symptoms of COVID-19 are fever, dry cough, and tiredness. Other symptoms that are less common and may affect some patients include aches and pains, nasal congestion, headache, conjunctivitis, sore throat, diarrhea, loss of taste or smell or a rash on skin or discoloration of fingers or toes. These symptoms are usually mild and begin gradually. Some people become infected but only have very mild symptoms [31]. These symptoms may be developed in patient within 14 days of infection of coronavirus. Higher mortality rate and delay in the identification of infection are the main challenges in major countries affected by a coronavirus. Due to the unavailability of vaccine or antiviral medicine to prevent coronavirus, detecting infection at an early stage based on symptoms and identifying the accurate treatment to recover the patients from infection can help to decrease the mortality rate and increase the recovery rate of patients from coronavirus.

The proposed approach contributes to solving the above challenges by collaborative healthcare data mining with data related to coronavirus patients. Major hospitals treating patients with coronavirus collect the patient's data like different symptoms, age, location, gender, treatments, etc. using the EHR system in hospital. Data mining solution of a disease requires the more patients' data for better accuracy. Hence, these hospitals with EHR systems collaborate to perform data mining on global patients data from all collaborative hospitals.

The proposed approach helps to collaborate healthcare organizations for combating COVID-19 using global data mining solutions with data from all healthcare organizations while preserving the privacy of healthcare data. Physicians and medical researchers can take advantage of this global healthcare mining results for combating COVID-19.

Conclusion

In this paper, we focused on improving the accuracy of disease identification using the collaboration of multiple EHR systems on horizontally partitioned data. Our proposed approach computes high accuracy association rules from the distributed data available at multiple EHR systems while preserving privacy. Experimental results using breast cancer data show the benefits of the proposed approach compared to results of any single EHR system. The proposed approach helps to identify the disease at an early stage using accurate association rules generated from distributed healthcare data. This approach can also be applied for early identification of COVID-19 infection in patients using some symptoms of patients.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Nahar J, Imam T, Tickle KS, Chen YPP. Association rule mining to detect factors which contribute to heart disease in males and females. Expert Syst Appl. 2013;40(4):1086–93.
2. Covid-19 coronavirus pandemic. [Online] 2020. https://www.worldometers.info/coronavirus/. Accessed 15 May 2020.

3. Heron M. Deaths: leading causes for 2015. Natl Vital Stat Rep. 2017;66(5):1–76.
4. Tang PC, McDonald CJ. Electronic health record systems. Biomed Inform. 2006;10(4):447–75.
5. Ordonez C. Association rule discovery with the train and test approach for heart disease prediction. IEEE Trans Inf Technol Biomed. 2006;10(2):334–43.
6. Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. Int J Healthc Biomed Res. 2013;1:94–101.
7. Shin AM, Lee IH, Lee GH, Park HJ, Park HS, Yoon KI, Lee JJ, Kim YN. Diagnostic analysis of patients with essential hypertension using association rule mining. Healthc Inform Res. 2010;16(2):77–81.
8. Karabatak M, Ince MC. An expert system for detection of breast cancer based on association rules and neural network. Expert Syst Appl. 2009;36(2):3465–9.
9. Koh HC, Tan G, et al. Data mining applications in healthcare. J Healthc Inf Manag. 2011;19(2):65.
10. Clifton C, Kantarcioglu M, Vaidya J. Defining privacy for data mining. In: National science foundation workshop on next generation data mining, vol 1. 2002; pp. 199–204.
11. Gkoulalas-Divanis A, Loukides G, Sun J. Publishing data from electronic health records while preserving privacy: a survey of algorithms. J Biomed Inform. 2014;50:4–19.
12. Kantarcioglu M, Clifton C. Privacy preserving distributed mining of association rules on horizontally partitioned data. IEEE Trans Knowl Data Eng. 2004;9:1026–37.
13. Bogetoft P, Christensen DL. Secure multiparty computation goes live. In: Proceedings of financial cryptography and data security. Springer; 2009. pp. 325–43.
14. Hirt M, Maurer U, Przydatek B. Efficient secure multi party computation. In: Advances in cryptology. Berlin: Springer; 2000. p. 143–61.
15. Moses TJ, Elavarasi K, Jayachitra J. Privacy preserving mining of association rules in horizontally distributed databases. Int J Manag IT Eng. 2014;4(9):209–22.
16. Ouda MA, Salem SA, Ali IA, Saad ESM. Privacy-preserving data mining (ppdm) method for horizontally partitioned data. Int J Comput Sci Issues. 2012;9(5):339–47.
17. Patel AC, Rao UP, Patee DR. Privacy preserving association rules in unsecured distributed environment using cryptography. In: The proceedings of the third international conference on computing communication and networking technologies (ICCCNT-12); 2012. pp. 1–5.
18. Kantarcioglu M. A survey of privacy-preserving methods across horizontally partitioned data. In: Proceedings of privacy-preserving data mining of advances in database systems. Springer; 2008. pp. 313–35.
19. Pedersen TB, Saygın Y, Savaş E. Secret sharing vs. encryption based techniques for privacy preserving data mining. In: Eurostat work session on statistical disclosure control. Eurostat; 2007.
20. Modi CN, Rao UP, DR. Elliptic curve cryptography based mining of privacy preserving association rules in unsecured distributed environment. In: Proceedings of IEEE international conference on advances in communication, network and computing; 2010. pp. 94–8.
21. Ge X, Yan L, Zhu J, Shi W. Privacy preserving distributed association rule mining based on the secret sharing technique. In: Proceedings of 2nd international conference on software engineering and data mining (SEDM). IEEE; 2010. pp. 345–50.
22. Nanavati NR, Jinwala DC. A novel privacy-preserving scheme for collaborative frequent itemset mining across vertically partitioned data. Secur Commun Netw. 2015;8(18):4407–20. https://doi.org/10.1002/sec.1377.
23. Nanavati NR, Lalwani P, Jinwala DC. Analysis and evaluation of schemes for secure sum in collaborative frequent itemset mining across horizontally partitioned data. J Eng. 2014;2014:110–20.
24. Chahar H, Keshavamurthy BN, Modi C. Privacy-preserving distributed mining of association rules using elliptic-curve cryptosystem and Shamir's secret sharing scheme. Sādhanā. 2017;42(12):1997–2007.
25. Galbraith SD. Elliptic curve Paillier schemes. J Cryptol. 2002;15(2):129–38.
26. Shamir A. How to share a secret. Commun ACM. 1979;22(11):612–3.
27. Burdick D, Calimlim M, Gehrke J. Mafia: a maximal frequent itemset algorithm for transactional databases. In: Proceedings 17th international conference on data engineering. IEEE; 2001. pp. 443–52.
28. Hussein M, El-Sisi A, Ismail N. Fast cryptographic privacy preserving association rules mining on distributed homogenous data base. In: Knowledge-based intelligent information and engineering systems. Berlin: Springer; 2008. p. 607–16.
29. Nguyen XC, Le, HB, Cao TA. 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.
30. Breast cancer wisconsin (original) data set. [Online]. https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data. Accessed 28 Jan 2020.
31. Q/a on coronaviruses (covid-19). [Online] 2020. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses:text=symptoms. Accessed 2020.