IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 6, 2021, accepted January 15, 2021, date of publication January 19, 2021, date of current version January 27, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3052799

An Efficient Association Rule Mining From Distributed Medical Databases for Predicting Heart Diseases

AHMED M. KHEDR[1,2], ZAHER AL AGHBARI[1], (Senior Member, IEEE), AMAL AL ALI[3], AND MARIAM ELJAMIL[4]

[1]Department of Computer Science, University of Sharjah, Sharjah 27272, United Arab Emirates
[2]Department of Mathematics, Zagazig University, Zagazig 44519, Egypt
[3]Department of Information Systems, University of Sharjah, Sharjah 27272, United Arab Emirates
[4]Department of Computer Science, Ain Shams University, Cairo 11566, Egypt

Corresponding author: Ahmed M. Khedr (akhedr@sharjah.ac.ae)

ABSTRACT Electronic Health Records (EHRs) are aggregated, combined and analyzed for suitable treatment planning and safe therapeutic procedures of patients. Integrated EHRs facilitate the examination, diagnosis and treatment of diseases. However, the existing EHRs models are centralized. There are several obstacles that limit the proliferation of centralized EHRs, such as data size, privacy and data ownership consideration. In this paper, we propose a novel methodology and algorithm to handle the mining of distributed medical data sources at different sites (hospitals and clinics) using Association Rules. These medical data resources cannot be moved to other network sites. Therefore, the desired global computation must be decomposed into local computations to match the distribution of data across the network. The capability to decompose computations must be general enough to handle different distributions of data and different participating nodes in each instance of the global computation. In the proposed methodology, each distributed data source is represented by an agent. The global association rule computation is then performed by the agent either exchanging some minimal summaries with other agents or travelling to all the sites and performing local tasks that can be done at each local site. The objective is to perform global tasks with a minimum of communication or travel by participating agents across the network, this will preserve the privacy and the security of the local data. The proposed association rule mining methodology will be used for heart disease prediction using real heart disease data. These real data exist at different clinics and cannot be moved to a central site. The proposed model protects the patient data privacy and achieves the same results as if the data are moved and joined at a central site. We also validate the extracted association rules from all the data providers using an independent test datasets.

INDEX TERMS Association rules, medical distributed databases, electronic health care records, agents, data privacy.

I. INTRODUCTION

Electronic Health Care Records (EHRs) are widely used in various healthcare institutions, such as hospitals and clinics. EHR contains a great collection of medical and health care information on symptoms, diagnosis, laboratory tests, other diagnostic tests, and treatments and is therefore a potential source of data for pragmatic trials, evaluations of drug safety, epidemiologic studies, and various health care organization evaluations. Information recorded in the EHRs has the capability to revolutionize a health care system through careful analysis and interpretation of data, feedback, and change implementation. Their utilities are not limited to supporting financial and administrative operations or maintaining patient records.

The use of EHRs also help to deliver efficient healthcare and improve the quality of patient care. Careful evaluation of EHR data quality and applicability can support different scientific and clinical evaluations. With the right mining approach, EHRs can be an effective tool for diagnosing and predicting disease. It can also help in prescribing treatments [1], [2]. This will help to provide

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Huo.

international standards for interoperable applications that use health, economic, social, behavioral, and environmental data to communicate, interpret, and act intelligently upon complex healthcare information to promote precision medicine and a learning health system. In the past years, various data mining algorithms have been developed to extract data from the EHR [1], [3]–[6].

A. DISTRIBUTED DATA SOURCES

Computing situations that are beginning to emerge in the networked environment require data and knowledge from a number of geographically distributed sites to be simultaneously considered. A number of geographically distributed databases together form an implicitly specified global dataset that contains all the data relevant for a computation. For example, some pattern discovery tasks may require simultaneous consideration of data, parts of which reside in medical databases, labor statistics databases, and employment related databases. Each of these is a huge database and resides on a different site in a different city (clinics or hospital). One cannot hope to easily move all these databases to a single computer site, merge or join them, and then execute an algorithm with the tuples in the resulting humongous database.

It would be desirable to have algorithms that let the individual databases reside at their own sites and work with an imagined implicit join of the databases by decomposing themselves into localized computations such that each localized computation can be performed locally within a single site using its physical database. A common constraint in these situations is that the data cannot be moved to other network sites due to security, size, privacy and data ownership considerations. An example of such situations is, we may need to compute decision trees, association rules, or some complex statistical quantities using data from a census database, a diseases database, a labor statistics database, and a few pollution databases located in ten different cities across the country. It is impossible to bring these databases together and join them for performing some computations. Also, a new instance of some computation may require data from a different set of participating nodes and databases. In real life, medical data is naturally distributed. That is, each database is distributed and owned by an individual organization (hospital), where data from various organizations, or hospitals, cannot be moved due to privacy, size, or data ownership considerations.

In this paper, we propose a new approach to apply the association rules algorithm in a distributed fashion for predicting heart diseases, without the need to move the local datasets. Some data mining projects may require data from multiple sites. In order to enable shared access to the healthcare record, data integration is necessary [7]. The obstacle we are dealing with is, the selection of essential data from these geographically distributed systems, without moving databases. By not moving local datasets, such distributed healthcare systems increased the privacy of patients' data. Another rising issue in healthcare systems, is the distribution of patient information among various healthcare organizations. Since each healthcare organization has only the records of their clients (patients), most hospitals are limited to constructing their individual decision-making models, with the help of minimally accessible data at their electronic healthcare systems. Having generous load of data is crucial for building and training decision models from medical input [5]. Therefore, we introduce a privacy-preserving data mining scheme for horizontally distributed medical data. We develop a data integration model using association rules for predicting heart diseases, given medical data from various health organizations, without moving or exposing patient records from any of the particular resources (i.e. hospital). Our objective is to set up a disease prediction model that can obtain data from various facilities, located at different geographical locations, without exposing patient's data.

B. DATA MINING

Data mining is the process of discovering interesting patterns and knowledge from large amount of data. The data sources can include databases, data warehouses, other information repositories, or data that are streamed into the system dynamically [8]. In healthcare systems, numerous medical data are stored in databases in the form of images, charts, texts and digits. Unfortunately, the surplus stored medical data is not widely used for disease diagnosis and prediction purposes. The question is, how to convert such stored data into information and knowledge? And the answer to making knowledge discovery achievable from the stored data is data mining tools.

Many research works emphasize collecting the right information, allowing correct decisions or assistance to be taken in the timely prevention of illness. The extraction of accurate knowledge from stored data not only helps to maintain and prevent certain diseases, but also helps to diagnose and treat patients. Association rules play an important role as a data mining technique, in the medical field, for boosting disease prediction. They can predict diseases from a few subsets of attributes. Association rules can also avoid dataset fragmentation because every rule may be overlapping with other rules. For these reasons, association rules are in a favorable position among standard supervised machine learning algorithms such as Support Vector Machine (SVM) [9], logistic regression [9] and Decision Tree [10], [11]. In medical domain, a concern that may arise while using association rules is the massive number of produced rules due to low support metric. Consequently, a large number of medically irrelevant rules are included in the produced set. For these reasons, search constraints were introduced in [4], [12] to eliminate irrelevant patterns. These also reduce data overfit while searching for association rules, which are then validated using a separate dataset (test sample) [9].

The association rules were generated from medical heart datasets with heart perfusion measurements in [3]. Each rule serves as an anticipating motif, and is part of a larger dataset. From a medical perspective, the association rules provide associations of individual sets of attributes; for example, heart

measurements and risk factors. When using binary attributes, the aim is to find the existence or absence of a heart disease. Recently, privacy preserving techniques for horizontally and vertically distributed systems, were established on a star topology where an agent is required [7], [13]–[17]. Every facility distributed data with an untrusted third party-central agent. This agent was responsible for establishing the integrated predictive design. However, these models were subject to high communication costs due to frequent data transfers with the central agent. Therefore, in our approach, we are designing a new prediction model, with information gathered from various horizontally distributed databases, without transferring confidential patient records. As a result, patients' privacy is preserved. Our method introduces distributed algorithm for mining association rules, based on weighted rules which are generated in accordance with rule's confidence and local data size. Then, the integrated model uses an independent test data to generalize the extracted rules on the distributed databases, instead of having specific rules for each individual local database. The main contributions of our work are summarized as follows:

1) Design a model to deal with the distributed databases, based on constrained association rules and minimum metrics, for horizontally distributed systems. Search constraints are used to eliminate the number of irrelevant patterns produced.
2) Design an integration algorithm that does not need a third-party agent. The integration process generalizes the rules of the extracted association by obtaining a medical summary only from individual healthcare organizations, without affecting the privacy of patient records.
3) Propose a validation process using the train and test approach. We validate the integration of the predictive decision-making model using separate (disjoint) test dataset.
4) Demonstrate the performance of the proposed decomposable algorithm using heart-narrowing patient records, collected from various geographical locations.

The article is organized into various sections as follows: Section II presents the problem formulation and the integration methodology of the databases, Section III presents the related work, Section IV presents the proposed technique, Section V presents the experimentation and results, and finally, Section VI concludes the paper.

II. PROBLEM FORMULATION

In this section, we provide the description on different types of data distribution and the proposed methodology to manage these medical distributed databases without moving and join at one node. As mentioned earlier, we assume that each node possesses a component database containing a set of attributes.

Computerized data from healthcare institutions is expanding exponentially as a result of the use of new information systems. These information systems are segregated on the basis of the type of data stored and where they are used. They work separately from each other, because they are localized stand-alone systems. In order to improve the quality of healthcare services by granting shared access to data, healthcare systems are opting for the integration of data sources. Hospitals call for special integration architecture for several purposes, such as avoiding data redundancy, which may result in extra work and extra communication costs, and supporting the decision-making process that may help in diagnosis and prediction of disease. In addition to previous applications, patient data are available for disease treatment, administration and research purposes. Data privacy and ownership should be considered before embracing EHRs systems [5], [18], [19].

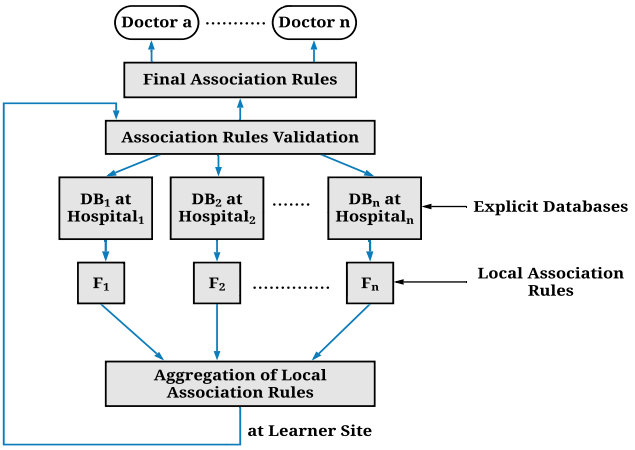

FIGURE 1. Model diagram: EHRs.

In this work, we consider that each medical database component resides at different local sites (clinics or hospital). Fig. 1 demonstrates the model of a distributed EHRs (various local sites such as pharmacies, hospitals and clinics) and the integrated model. Each medical source is geographically positioned at a distinct location. Moving datasets from local sites to another location is not a valid solution for a variety of reasons, such as data ownership consideration, privacy, communication overhead and data size. However, a generous amount of data is required to build a decision model from medical data.

Medical data are dispersed across various distributed databases; thus, shared access is crucial for practitioners (doctors, nurses and pharmacists) to deliver better patient care. In order to help practitioners utilize data mining tools and to provide easy access to efficient information, data integration is required for administrative and health implementation. In this paper, we are introducing a model to handle distributed databases and an association rule algorithm to mine this model.

In this research, patient data containing numerical and categorical values are transformed into transactions where each item relates to a single numeric domain or categorical value. The integration methodology of the databases is as detailed below. Table (1) shows the symbols used in this paper.

TABLE 1. Notations used.

Notations	Description
ψ, α, λ	support, confidence, and lift respectively
τ	training fraction
D	global database
d_i	local database at i^{th} site
M	medical records
S	transformed medical records
R_{train}	rules generated from S_{train} at global site D
r_{train}	rules generated from S_{train} at local site d_i
R_{test}	rules generated from S_{test} at global site D
r_{test}	rules generated from S_{test} at local site d_i
R_P, R_A	rules for prediction, (presence) and (absence) respectively

A. IMPLICIT GLOBAL DATABASE

The implicit global database D exists as fragments that are distributed over various local sites. Each site s_i stores a component D_i of the database D, containing data at s_i in the form of tuples such that each tuple includes a distinct or same set of attributes. The distribution strategy of databases require ''move and Join'' operation to construct the implicit global database D from the components D_i's.

B. INTEGRATION OF DISTRIBUTED DATABASES

In this section, we assume that the global database D is distributed as local database components over all the local sites The distribution described above is considered as an implicit global database. The global database D can be generated at the end user or central site through the join of its component relations and can provide remarkable data suitable for performing computation as well as mining activities using association rules. The key focus of the proposed scheme is to mine the implicit global database D using association rules by maintaining the data locally as D_i at the local site s_i and minimize the data communication between the sites. As a result, the local results of each database D_i at a site s_i will only be transmitted to the coordinator or central site for aggregation.

The mathematical formulation of the proposed problem can be described as follows: consider for example, the set of n components of databases D_i's are distributed at sites s_i's where, each site holds a set of tuples. Then, the global implicit D includes the union of tuples of the local components, as shown in Equation 1.

$$D = \bigcup_{i=1}^{n} D_i, \tag{1}$$

The proposed scheme emphasizes on determining the association rules of implicit D through minimal communication of messages among the local sites. Therefore, the global computation task is divided into local computations. As a result, the aggregation of local computation results can help to produce the global association rules. This can be formulated mathematically in general as follows: Consider that a function F is applied on the database D to obtain the result

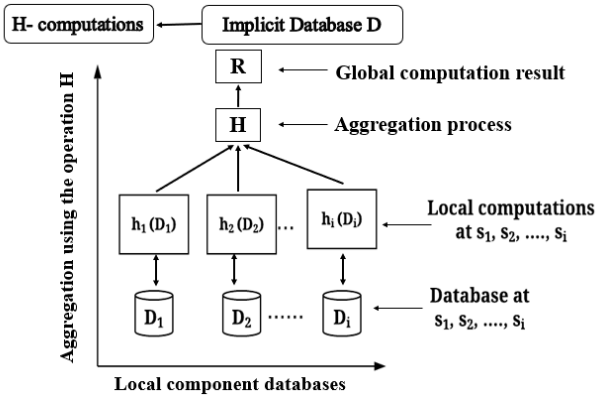

FIGURE 2. Aggregation process at coordinator site using statistical summaries from the local sites.

R as given in Equation (2). As stated previously, the required distributed computation is the derivation of association rules corresponding to D. Here, we can denote F as the algorithmic implementation for the derivation of association rules for D, where R represents the obtained association rules for D.

$$R = F(D), \tag{2}$$

Each local site performs local computation on its respective database component. The local results are then exchanged through communication with the central site (or coordinator) to obtain the global results of computation from the components D_1 to D_n. The corresponding realization of function F as given in Equation (2) can be rewritten as:

$$R = H[h_1(D_1), h_2(D_2), h_3(D_3) \ldots, h4_n(D_n)], \tag{3}$$

Here, $h_i(D_i)$ represents the i^{th} s_i's local computation implemented on D_i at s_i such that, the operation H represents the aggregation of the results of local computations executed by the coordinator. Each problem involves a distinct set of h-operators (h_is) and the features of H and h_i relies on the participated D_is. Finally, the central site would compute R from the D_is. Fig. 2 shows the process by which the coordinator site computes R from the D_is. A local computation $h_i(D_i)$ is performed at every local site s_i using the database D_i. The results of these local computations are aggregated using the operation H at the coordinator.

III. RELATED RESEARCH

The purpose of using Electronic Health Records (EHRs) is to provide more efficient and effective health care [20]. As health care systems are expanding, digital data is also increasing. These systems are grouped based on their location, and the sort of data stored in them; such as Hospital Information Systems (HIS), Radiology Information Systems (RIS), Laboratory Information Systems (LIS) and Picture Archiving & Communication System (PACS) [21]. Even with the advantages of EHRs in disease diagnosis and prediction, they are still not widely adopted by physicians [22]. Some of the barriers that restrain physicians from using EHRs

are concerns about confidence and privacy [23], inability to find an EHRs that meets their needs [24], and lack of data exchange between EHRs and other data systems such as labs, oncology, and cardiology [25]. The main barrier may lie in inter institutional integration, where practitioners are unable to access EHRs to benefit from patient data, and have to refer hard copy documents [26].

Data mining is used in many healthcare applications by extracting numerous, useful information concealed in huge quantities, and providing decision support [2], [27]. The study in [28] gives a general view for data mining process in healthcare, by giving descriptive information on knowledge discovery in databases, data warehousing, Business Intelligence and data mining methods. Patient's old health records, test history, and individual details are utilized by data mining tools. Thus, precautions can be taken if symptoms are detected, and physicians can minimize the effect or even avoid the disease [29]. The use of data mining in healthcare provides numerous advantages such as, recognizing fraud in health insurance, prediction and diagnosis of disease, hospital infection control, and building smart treatment system with recommended drugs suitable for patient's case [30], [46]. In [6] the author provides a study of data mining and knowledge discovery tools in healthcare. It examines the use of different data mining tools such as classification, clustering, association, and regression. For example, artificial neural network (ANN) is used in prediction of skin cancer using a multi-parameterized artificial neural network to predict the risk of growing non-melanoma skin cancer in [31]. Also, convolutional neural network (CNN) effectively predict chronic disease eruption in disease communities, and was examined on cerebral infarction disease in [32]. Despite its advantages, ANN requires high execution time and it is hard to explain.

Classification technique such as k-nearest neighbor (KNN) was used in [33] to predict the chronic kidney disease from health data with environmental elements. Unlike neural networks (NN), KNN is simple to be implemented and trained, however, it is sensitive to noise. Another classification technique is decision tree (DT), which is easy to interpret and can handle both numerical and categorical data. However, it is limited for a single output attribute and its performance relies on dataset type, which makes it an unstable technique. Several decision trees were used in [34] to predict liver diseases. The paper provides a comparison of calculated performance of different decision tree models, such as e J48, Logic Model Tree (LMT), Random Forest, Random tree, Reduced Error Pruning tree (REPTree), Decision Stump, and Hoeffding Tree. In other studies, multiple classification techniques were used to improve prediction and classification, such as combining decision tree with neural network in [35], [36]. In these studies, DT and NN were used to predict lung tumor and heart attack, respectively. Support vector machine (SVM) is one of the most used classifiers by researchers in the medical domain because of its high accuracy compared to other classifiers. This is because SVM classifiers have less over fitting issues and can simply deal with complex nonlinear data points. SVM is used in various healthcare studies, such as genetic SVM technique in analyzing heart valve disease. The model in [37] selects essential features, and classify the output signal from ultrasound of heart valve. In [38], [39], SVM was used to evaluate the classification of malignant and benign breast cancer. However, SVM is computationally expensive, as it requires accurate kernel function and to produce distinct output for each dataset.

Association rules (ARs) is an important data mining method in the medical field as it searches for interesting patterns and relationships between diseases and symptoms. The use of association rules has been extended to develop predictive models [40]. Association rules were used to find associations between perfusion measurement attributes of activated brain areas for early diagnosis of Alzheimer's disease in [41]. The process started by searching for activated brain regions of interest using estimation methods, which are then used as an input for association rules algorithm. The proportion of activated regions of the brain are related to support and confidence metrics. Using association rules with different classifiers, results in higher output accuracy and better quality, compared to the individual classifiers as shown in [42], which proposes a predictive model for Dengue fever using multiple rule-based classifiers (decision tree, rough set classifier, naive Bayes, and associative classifier). Performance of each classifiers separately, as well as all classifiers combined were discussed. In [49], a fast association rule mining algorithm with low memory requirements based on the MapReduce framework was proposed. The work in [50] defined a new concept that produces multitask rules.

All of the above works use sequential algorithms and are not suitable for dealing with distributed medical databases, due to efficiency and privacy concerns. Obligations to improve efficiency in healthcare facilities, have stimulated studies for the integration of medical data sources that are naturally distributed [18]. Several database integration methods have been introduced in [43] for around 900 databases. The distinction between biological and clinical data has been taken into account during the process. Other methods used to integrate heterogeneous databases are grid technology, ensemble approaches [5] and association rules [44]. Many privacy-preserving approaches for mining horizontally distributed databases were proposed, such as support vector machine [18], naive Bayes classifier [45] and ensemble-based classifiers [29]. The previous models have drawbacks such as higher computation cost due to repeated contact with the central agent, however, in our distributed model, each local site performs additional computation, and then the only the information summary will be transferred to the coordinator site, which means that less information needs to be transmitted. The importance of this lies in the fact that the cost of communication is much higher than the cost of computing. Moreover, only summaries, which boosts the privacy of the medical records. In [37], a model was proposed to study the association of blood pressure with the risk of cardiovascular disease, in future. The problem with the model was that the

study population was collected from Korea only; therefore, the results cannot be generalized to different races or ethnicities.

In our model, we avoid this drawback in association mining model as follows:

- We apply our model on naturally horizontal distributed databases using data collected from various countries.
- We validate association rules generated at each local site using training dataset, which then results in generalizable rules for all local databases.

IV. EXTRACTING ASSOCIATION RULES FROM DISTRIBUTED MEDICAL DATA

An effective assessment approach for evaluating data mining prototypes is by dividing data into training and testing datasets. We first create a predictive model using a training dataset and then validate it using a separate test dataset. The objective is to minimize the effects of data discrepancies, provide better generalization when using new data, and improve the model accuracy. First, we control the size of each set by setting a training fraction, then we split S into S_{train} and S_{test}. Both S_{train} and S_{test} have minimum metrics of support, confidence and lift.

$$S = S_{train} \cup S_{test}, \ and \ S_{train} \cap S_{test} = \phi, \tag{4}$$

Given database D that consist of a set of transactions T, where a transaction t is a set of itemsets. Consider item sets $X, Y; X \implies Y$ is an association rule. The support of X with respect to T is defined as the proportion of transaction t in dataset which contains the itemset X.

$$supp(X) = \frac{|\{t \in T;\ X \subseteq t\}|}{|T|}, \tag{5}$$

The confidence value of a rule $X \implies Y$, with respect to a set of transactions T, is the proportion of the transactions that contains X, which also contains Y.

$$conf(X \implies Y) = \frac{supp(X \cup Y)}{supp(X)}, \tag{6}$$

The lift is the ratio of the observed support to that expected if X and Y were independent.

$$lift(X \implies Y) = \frac{supp(X \cup Y)}{supp(X) * supp(Y)}, \tag{7}$$

Our decomposable solution is splitted into three algorithms: The first one (Algorithm 1) will be executed by the coordinator site, it aggregates and returns the received results from the participating sites. The second algorithm is the Training algorithm (Algorithm 2) which will be executed by participating sites and returns the local association rules associated with their support, confidence and the size of the their databases. The third one is the Testing algorithm (Algorithm 3) which will be executed by the participating sites and its goals are to test the aggregated association rules by the coordinator site and returns the result of testing to the coordinator site. The training phase includes: (1) Medical Dataset Transformation, (2) Constraints for Association Rules, (3) Integration of EHRs. The test phase takes as input the integrated association rules, validates the rules on the test dataset and then produces general association rules. Both training and test datasets have a set of six measurements. The first set is used to search for motives (frequent item sets) during the training process. The second set is used to validate the recognized motives during the testing phase.

In training phase: For each local site D_i, Medical records M are first transformed into S itemsets. Training and test samples are split based on training fraction. In phase 1 of training phase, we generate k-itemsets based on association size, and group constraint. We define our predictive goal prior to items generation. In phase 2 of training phase, we remove irrelevant rules based on antecedent-consequence (AC) constraint. After searching for constrained association rules in each local database D_i, we integrate association rules at coordinator site D. We then use weighted rules from R_{train} to build the integration model. Weighted rules are obtained by measuring the confidence of each rule. If confidence is above a threshold, we use Equation (8) to calculate the local weight of the rule. The size of a local site is determined by the number of tuples at site i. We then calculate the global weight of the rule using Equation (9) based on domain expert threshold determined by medical experts; a rule is eliminated if its weight is below a minimum value.

In testing phase: We compute some metrics on S_{test} to validate the rules by removing specific rules with confidence level below the minimum threshold. The set of predictive rules, IR, is produced as a subset validated rules on test dataset. It also validates integration process using test data. Algorithm (1) gives the decomposable algorithm to find predictive rules for horizontally distributed systems. The details of each phase is as explained below:

A. TRAINING PHASE PROCEDURE

The procedure will be executed by every local site on its database (Training Data), see Algorithm 2.

1) Medical Data Set Transformation:
   To simplify the problem, attributes are treated as numeric or categorical data. Given a patient implicit data D with N tuples, where $D = a_1, a_2, \ldots, a_N$, with numeric, categorical, D is converted to a transactional dataset $S = S_1, S_2, \ldots, S_N$. We assign an item to each categorical value to transform attributes which are categorical to items, while numeric attributes are discretized (binned) and each bin becomes an item.
2) Constraints for Association Rules:
   To find predictive association rules and eliminate irrelevant motifs (frequent item sets), search constraints are introduced. In order to discover association rules, we divide the problem into two steps: Step-1 searches for associations k-itemset X such that support $\psi_i(X) \geq \psi$, and Step-2 searches for all rules $X \implies Y$ s.t. confidence $\alpha_{r_i}(X \implies Y) \geq \alpha$ and lift $\lambda_{r_i}(X \implies Y) \geq \lambda$. Four constraints are defined, three are used in Step 1 and one in Step 2. Item filtering constraint,

Algorithm 1 Distributed Algorithm to Find Predictive Rules for Distributed Databases- Coordinator Site

1: **Input:** Global implicit distributed database D among N sites.
2: **Output:** Integrated Predictive Rules, IR.
3: Find predictive rules, R_{train_i}, for each local database D_i at site i by calling $Training()$
4: Integrate the received predictive rules: $R_{train_1}, R_{train_2}, \ldots, R_{train_N}$.
5: Compute the weighted rules R_{train} as follows:
6: compute total number of tuples ($N_{total} = \sum_{i=1}^{N} N_i$, where N_i is the number of tuples at site i.
7: For each rule r in R_{train}, compute:
8: local weight, Lw:

$$Lw(r) = \alpha_i * N_i, \tag{8}$$

where α_i is the confidence of a rule r at site i and N_i is the size of a local dataset D_i at site i.
9: global weight, Gw:

$$Gw(r) = \frac{\sum_{i=1}^{N} Lw(r)}{N_{total}}, \tag{9}$$

10: Eliminate rules if $Gw_i < \epsilon$, where ϵ is the minimum global weight threshold.
11: Validate integrated R_{train} on S_{test_i} at each site:
$R_{test_i} = Testing(R_{train}, D_i)$
12: From the received R_{test_i}, eliminate a rule $r_i \in R_{test_i}$ if $\psi_i(X \implies Y) < \psi$ or $\alpha_i(X \implies Y) < \alpha$.
13: Finally, set $IR = R_{test} = R_{test_1} \cup R_{test_2} \cup \ldots \cup R_{test_n}$

Algorithm 2 Distributed Algorithm to Find Predictive Rules for Distributed Databases- Training Phase Procedure

Divide your database D_i into S_{train} and S_{test} based on λ.
2: Apply Item filtration constraint before generating itemsets based on predictive goal, G("Presence") or G("Absence").
Using minimum support and the group() constraint, find for frequent k-itemsets on S_{train} for $k\{1, \ldots, k\}$.
4: Produce association rules r_{train_i} using ac() constraint, confidence and lift thresholds.
Send back to the coordinator site the produced association r_{train_i} rules with support, confidence for each rule and number of tuples at database D_i (N_i).

Algorithm 3 Distributed Algorithm to Find Predictive Rules for Distributed Databases- Testing Phase Procedure

Validate the received integrated rules R_{train} on S_{test}.
2: For each frequent itemset X, compute test support $\psi_i(X, S_{test})$.
For each rule $X \implies Y \in r_{test}$, compute test support $\psi_{r_i}(X \implies Y)$,
4: Compute test confidence $\alpha_{r_i}(X \implies Y)$ on S_{test}.
Send back to the coordinator site the tested produced association rules R_{test} that have confidence greater than threshold with their support, confidence, and number of tuples of the database.

rule size constraint, and group constraints are used in the first step, while antecedent consequent constraint is applied in the second step.

Step 1: Depending on the prediction target G, item filtering constraint is defined. There are item sets to predict the presence of disease (G = "Presence") and some for the nonexistence of the disease (G = "Absence"), while some items are utilized for both predictions. Item filtering constraint is applied prior to the generation of item set [47]. The rule size constraint is the second constraint to be applied in phase 1 during the generation of item sets. We generate frequent item sets up to size k, by eliminating frequent item sets of size $k+1, k+2, \ldots$ etc. We then use group constraint to prevent association between items related to risk factors and items related to heart measurements. Risk factors and heart measurements are mentioned in Medical Data Description in section V.

Step 2: The antecedent-consequent, ac, constraint will now be applied to the rules instead of associations. (AC) is a sample of predictive rules and defined as $ac(A_j) = C_j$. We define a set of AC constraints on $\{A_1, \ldots, A_p\}$ by $C = \{C_1, \ldots, C_p\}$. $C_j = 1$ if attribute A_j emerges exclusively in the antecedent of a rule; $C_j = 2$ if A_j emerges exclusively in the consequent.

3) Integration of EHRs:
The data being computerized by healthcare institutions has led to exponential expansion, due to new information systems being employed. These information systems are segregated on the basis of type of data stored and where they are being used. These systems work independently, because they are localized stand-alone systems. Health data are naturally spread across different sites where each site produces its own association rules to integrate the association rules of these different sites. We first weigh the constrained rules resulting from the training sets r_{train} of each local site using the minimum confidence and database size of the rule r_i:

$$Lw(r_i) = \alpha_{r_i}(X \implies Y) * N_i, \tag{10}$$

where N_i is the number of tuples in D_i at site i. Then, we compute the total weight of each rule in R_{train} can be computed as follows:

$$Gw(r_i) = \frac{\sum_{i=1}^{N} Lw(r_i)}{N_{total}}, \tag{11}$$

where N_{total} is the total number of tuples in the implicit database D. In order to increase the efficiency of medical data, healthcare systems are integrating distributed dataset and thus, benefit from the diversity of the local datasets. The integration of medical databases is an

important topic and necessary for knowledge discovery. Hospitals may have particular features that requires special integration architecture. For example, privacy is a sensitive issue that should be thought of, during the integration process. In our model, we preserve the privacy of patients' records by sending only the data summary using association rules, in order to predict heart disease without exposing private information about the patients.

B. TESTING PHASE

The procedure will be executed by every local site on its database (Testing Data), see Algorithm 3. The integrated association rules are the input for testing phase. For each local site, the rules generated at the global site R_{train} are computed on training set S_{train}, and the rules $R_{test} \subseteq R_{train}$, such that R_{test} has minimum threshold measurements ψ, α, λ on test set S_{test}. R_{test} is computed by first setting $R_{test} = R_{train}$. Then, we obtain R_{train} by inspecting the association rules on S_{train} with respect to minimum metrics ψ, α, λ. Next, for every rule in R_{test}, confidence, support and lift are calculated in accordance with S_{test} and rules with measurements less than thresholds are eliminated from R_{test}.

C. PREDICTION OF A HEART DISEASE

Based on the value of k, each local site produces a number of frequent itemsets. As the value of k increases, the number of produced frequent itemsets increases and vise versa. To ensure correctness of the prediction of heart diseases, constraints are apply on the dataset to remove irrelevant data. That is the data that will contribute negatively to the result is removed. There are three types of constraints: *item filteration*, *group* and *ac*.

The *item filtration* constraint is used to keep, or remove, risk factors and heart measurements attributes from the dataset. The *group* constraint is used to keep, or remove, the heart measurements measurements attributes. The *ac*, which is antecedent consequent constraint enforces the rules to have only the *Num* attribute in the consequent, while rest of the attributes in the antecedents.

These constraints are applied on the data to produce association rules at each local site. Then, these rules R_{train} will be integrated and generalized for all sites. Next, we validate these rules using a test set S_{test} to generate R_{test}.

To predict heart diseases according to heart measurements and risk factors attributes, association rules are categorized into:

- Association rules predicting the absence of heart arteries narrowing.
- Association rules predicting presence of heart arteries narrowing.

A detailed discussion on the results of predicting heart diseases from association is in Section V.C.

TABLE 2. Heart-disease attributes (constraints: on = 1, off = 0).

Attribute Name	Medical Meaning	Constraints		
		item filter	group	ac
age	Patient age	0	0	1
sex	Patient gender	0	0	1
cp	Chest pain type	0	1	1
trestbps	Resting blood pressure	1	0	1
chol	Cholesterol	1	0	1
fbs	fasting blood sugar	1	1	1
restecg	resting electrocardiographic results	0	0	1
thalach	maximum heart rate achieved	0	0	1
exang	exercise induced angina	1	1	1
ca	number of major vessels	0	0	1
thal	inherited blood disorder	0	0	1
slope	slope of the peak exercise ST segment	0	0	1
oldpeak	ST depression induced by exercise	0	0	1
num	diagnosis of heart disease	0	0	1

V. EXPERIMENTATION AND RESULTS

We used Python to implement our algorithm and to obtain predictive rules for heart disease. The Algorithm (1) in Section IV is applied to the medical dataset shown in Table (2). We validated the results of our association rule generated from local sites by combining all local datasets into one central site and producing the association rule from that site. We noted that when the decomposable algorithm was executed, rules generated from local sites were the same as those generated from the central site.

This section is organized as follows: Section V-A gives the dataset description, section V-B includes dataset transformation and parameter settings, and section V-C provides the predictive rules produced from both training and testing phases with varying constraints and metrics, and final discussion.

A. MEDICAL DATASET DESCRIPTION

We used Heart Disease Records acquired from UCI Machine Learning Repository, combining numeric and categorical attributes. Some attributes like *sex* and *age*, *chol*, *fbs*, and *trestbps*, are treated as risk factor attributes. And other attributes like *exang*, *cp*, *restecg*, *thalach*, *slop*, and *oldpeak* are considered as heart indication measurements. Moreover, some attributes like *thal* shows the presence of inherited diseases. *Num* attribute indicates the target attribute for predicting disease. Table 2 shows the 14 attributes chosen for our study. Data were naturally distributed and collected from multiple institutions located at different geographical locations, such as Cleveland Clinic Foundation, Hungarian Institute of Cardiology, and V.A. Medical Center.

We selected almost 100 records from each institution to test the proposed decomposable distributed algorithm (Algorithm 1). In our experiments, therefore, $p = 14$ and $N = 300$.

B. PARAMETER SETTINGS

1) Transformation Parameters:
Table (2) displays the attribute names. We transform the values based on medical settings as follows: Most

TABLE 3. Number of associations varying k to predict presence of disease G("Presence") at distributed sites.

k	ψ	no. of associations (SITE 1)	no. of associations (SITE 2)	no. of associations (SITE 3)
3	0.01	3386	3488	3794
4	0.01	11214	10761	12217

TABLE 4. Constraints: number of rules at distributed sites (0 = off, 1 = on).

k	(SITE 1) G="Presence" constraints		(SITE 2) G="Presence" constraints		(SITE 3) G="Presence constraints	
	0	1	0	1	0	1
3	3523	97	3817	84	3794	86
4	23884	762	25635	701	29479	634

risk factor attribute values were transformed into binary values, and other heart measurements attributes were split into categories. *Trestbps* was split at cutoff 120 (normal) and 140 (high). Less than 120 was considered low. *Chol* was split at cutoff 200 (warning) and 250 (high). *Thal* was divided into categorical values according to type of inherited blood disorder, normal, fixed defect or reversible defect.

2) Metrics:
For training data, we used a training sample fraction of 60%. Minimum threshold support, ψ was set to 0.01. For rules to be medically credible, minimum confidence α is set to 70%, based on based on consultation of medical doctors. Lift threshold was adjusted to $\lambda = 1$.

C. RESULTS: PREDICTIVE ASSOCIATION RULES

1) Searching for Predictive Rules in Training Phase: Table (3) shows the impact of varying the maximum association size parameter k at each local site. k{3, 4} was used to search for frequent itemsets. As k increases, larger numbers of associations were produced.
Table (4) shows the impact of constraints on number of rules produced. Constraint = 0 illustrates that the constraint is off, and constraint = 1 illustrates that constraint is on. We divided the training sample into two sets, a set for predicting the presence of a disease, G = (''Presence''), and a set for predicting the absence of a disease, G = (''Absence''). Only the results for G = (''Presence'') were displayed, but both prediction goals had comparable results. We started by applying item filtration constraint before generating associations based on the prediction goal. When G = (''Presence''), risk factors indicating ''No'', and heart measurements indicating ''no warnings were filtered out. By using item filtration, the number of produced associations and rules were reduced.
The second constraint used was group constraint. For G = (''Presence''), heart measurements were the group constraints. While for G = (''Absence''), two group constraints existed, risk factors, and heart measurements. The third constraint was antecedent consequent constraint; only the *Num* attribute belonged to consequent, while rest of the attributes belonged to antecedents. Antecedent-consequent constraint does not change according to prediction goal. This constraint has a great effect on the number of produced rules, unlike group constraints which is not considered as effective. When constraints were ON, almost 97% of rules were filtered out. These are not the definitive rules, as validation process was to be applied. More rules were filtered out during the test phase and the final rules were obtained. These rules will not be specific to a local database, rather they will be generalized for the distributed database.

TABLE 5. Number of rules varying k to predict presence of disease G("Presence") at central site.

k	ψ	Central Integrated Site		Time
		train	test	
3	0.01	159	113	58
4	0.01	1691	697	194

TABLE 6. Number of associations and rules varying minimum support to predict presence of disease G("Presence") at global site.

K	ψ	no. of associations		no. of rules		Time
		train	test	train	test	
4	0.01	54463	21151	1691	697	194
4	0.1	4513	4209	236	118	46

2) Validating Rules Based on Training Set and Test Set: After the integration process, we validate the produced rules R_{train} using a test set S_{test} to generate R_{test}. This process was used to remove the specific rules and to test integration validity. In Table (5), when $k = 3$, 30% of rules are eliminated. This percentage increases into 60% when k is 4. This indicates the importance of validation process in reducing number of rules.
When varying minimum support in Table (6), the number of associations and rules were reduced by increasing threshold. By increasing the minimum support from 0.01 to 0.1, the numbers of associations and rules are decreased by about 80%. During the association process, only support metric is taken into consideration; while during rules generation, support, confidence and lift are considered. In Table 6, time indicates the execution time from transformation until validation. As shown in Table 5, time grows with larger k. It also increased by reducing support, which resulted in an increase in the number of associations.

3) Final Discussion on Predicting Heart Disease with Association Rules: Here, we elaborate on the importance of the medical rules in anticipating heart diseases according to heart measurements and risk factors attributes.
Rules are categorized into: (a) Association rules predicting the absence of heart arteries narrowing based on the absence of risk factors and normal measurements. (b) Association rules predicting presence of heart arteries narrowing based on the existence of risk factors and defect measurements. We search for rules with antecedents of $k = 4$, and high metrics. Parameters threshold are set as follows: $\psi = 0.01$, $\alpha = 70\%$ and

TABLE 7. Association rules predicting absence of heart disease G("Absence").

α = 1:
cp = typical, sex = M, $chol$(150, 220] $\longrightarrow$ num < 50 %diameter narrowing), ψ = 0.027, α = 1.0, λ = 1.57
cp = typical, $restbps \geq 140$, age(60, 100] $\longrightarrow$ num < 50 % diameter narrowing), ψ = 0.01, α = 1.0, λ = 1.57
sex = F, cp = non-anginal, fbs < 120) $\longrightarrow$ < 50% diameter narrowing), ψ= 0.13, α = 1.0, λ = 1.57
$\alpha \geq$ 0.9, $\lambda <$ 2:
age(40, 60], cp = atypical, sex = M $\longrightarrow$ num < 50% diameter narrowing, ψ = 0.1, α = 0.92, λ = 1.40
cp = non-anginal, age(40, 60], $restbps$ < 140 $\longrightarrow$ num < 50% diameter narrowing, ψ = 0.15, α = 0.9, λ = 1.48
cp = atypical, $restecg$ = normal, fbs < 120 $\longrightarrow$ num < 50% diameter narrowing, ψ = 0.109, α = 0.9, λ = 1.450
cp = atypical, sex = M, $restecg$ = normal $\longrightarrow$ num < 50% diameter narrowing, ψ= 0.08, α = 0.9, λ = 1.4
$thal$ = normal, $restbps$ < 140, $exang$ = 0 $\longrightarrow$ num < 50% diameter narrowing, ψ = 0.3, α = 0.9, λ = 1.4
$slope$ = upsloping, $exang$ = 0, $chol$ = (150, 220] $\longrightarrow$ num < 50% diameter narrowing, ψ = 0.127, α = 0.93, λ = 1.46
age(40, 60], cp = atypical, sex = M $\longrightarrow$ num < 50% diameter narrowing, ψ= 0.1, α = 0.91, λ = 1.401
α = 1, $\lambda <$ 2:
sex = F, cp = non-anginal, $chol$(250, 500] $\longrightarrow$ num < 50% diameter narrowing, ψ = 0.05, α = 1.0, λ = 1.57
sex = F, cp = non-anginal, age(60, 100] $\longrightarrow$ num < 50% diameter narrowing, ψ = 0.05, α = 1.0, λ = 1.5
cp = typical, $restbps$ geq 140, age(60, 100] $\longrightarrow$ num < 50% diameter narrowing, ψ = 0.018, α = 1.0, λ = 1.57
Two items (simple):
$thal$ = normal $\longrightarrow$ num < 50% diameter narrowing, ψ = 0.49, α = 0.85, λ = 1.34
$thalach$ = 179.0 $\longrightarrow$ num < 50% diameter narrowing, ψ= 0.01, α = 1.0, λ = 1.57
$oldpeak$ = 0.4 $\longrightarrow$ num < 50% diameter narrowing, ψ = 0.027, α = 1.0, λ = 1.57

$\lambda = 1$. Rules with minimum support are supposed to be available in at least three patients. Rules with confidence less than 70% are not medically reliable. Rules are selected after running distributed integration algorithm and after validation using test set.

a) Predicting Absence of Heart Disease: In Table (7), when confidence is 100%, rules with typical and non-angina chest pain appears with the absence of heart disease. There is a rule for male with low cholesterol level and another rule for an old age with blood pressure higher than 140. Both rules have low support. A rule for a female with fasting blood sugar less than 120 indicates absence of arteries narrowing.
We observe plentiful of chest pain in all rules, when confidence is higher than 90% with lift less than 2. There are important rules that show the link between absence of arteries narrowing and normal measurements. A rule with normal *thal*, resting blood pressure below 140, and false exercise induced angina appears with high support, as the rest of the rules. There are some rules for warning measurements, such as those for females with high cholesterol levels and females with old age, which still indicate lack of heart narrowing when combined with non-anginal chest pain. Another interesting rule is with high resting blood pressure and old age combined with typical chest pain. All of these rules appear when the support value is low, confidence is 100% and lift is less than 2. Lastly, simple rules with a single antecedent are generated. Each rule has either high support with low confidence and low lift, or low support with high confidence and high lift.
b) Predicting Presence of Heart Disease: As in rules predicting absence of a disease, in Table (8), chest pain attribute appears the most in presence of narrowing heart arteries as well. We selected high metrics rules with confidence 100% and lift higher than 2. Rules with high risk and defect measurements also appeared. For example, the presence of narrowing heart is associated with high resting blood pressure even in the absence of chest pain. Also, a defect *that* with the presence of *exang* are associated with greater narrowing of the heart arteries for older people. Interestingly, a fasting blood sugar that is higher than 120 is associated with the presence of a heart disease even with low support for people in the age between 40 and 60.
When confidence was between 90% and 100%, with lift above 2, and support value low, asymptomatic chest pain was found to associate with the narrowing of the heart arteries. These rules hold even in the presence of high *chol* level, and either flat *slope* or presence of *exang*.
Interestingly, when we dropped confidence below 90%, less items appeared in the antecedents. Narrowing of heart arteries is associated with defect *thal* or high *chol* level. These rules are associated with asymptomatic chest pain. Lift appeared strong in all rules. When restricting the rules to only two items, narrowing of heart arteries was associated with high rate of heart beat and exercise induced angina.

D. COMPLEXITY COMPUTING AND SECURITY CONSIDERATION

Traditionally, the algorithm complexity is evaluated in terms of memory and CPU time, however, this cost model is well-suited for computations on a single computer and the closely-coupled processors. When a number of loosely networked nodes are involved in a cooperative computation, the communication cost becomes the overwhelmingly dominant

TABLE 8. Association rules predicting presence of heart disease G("Presence").

α = 1, $\lambda \geq$ 2:
sex = F, $restbps \geq 140$, cp = asymptomatic $\longrightarrow$ num > 50% diameter narrowing, ψ = 0.018, α = 1.0, λ = 2.75
$thal$ = defect, $exang$ = 1, age(60, 100] $\longrightarrow$ num > 50% diameter narrowing, ψ = 0.027, α = 1.0, λ = 2.75
fbs > 120, age(40, 60], cp = asymptomatic $\longrightarrow$ num > 50% diameter narrowing, ψ = 0.03, α = 1.0, λ = 2.75
$\alpha \geq$ 0.9, $\lambda \geq$ 2:
$exang$ = 1, cp = asymptomatic, $chol$(250, 500] $\longrightarrow$ num > 50% diameter narrowing, ψ = 0.1, α = 0.91, λ = 2.5
$slope$ = flat, cp = asymptomatic, $chol$(250, 500] $\longrightarrow$ num > 50% diameter narrowing, ψ = 0.10, α = 0.9, λ = 2.5
$slope$ = flat, sex = M, cp = asymptomatic $\longrightarrow$ num > 50% diameter narrowing, ψ = 0.17, α = 0.9, λ = 2.48
$\alpha \leq$ 0.9, $\lambda \geq$ 2:
$thal$ = defect, cp = asymptomatic $\longrightarrow$ num > 50% diameter narrowing, ψ = 0.2, α = 0.8, λ = 2.32
cp = asymptomatic, $chol$(250, 500] $\longrightarrow$ num > 50% diameter narrowing, ψ = 0.13, α =0.78, λ = 2.1
$thal$ = defect, $exang$ = 1 $\longrightarrow$ num > 50% diameter narrowing, ψ = 0.17, α =0.8, λ = 2.3
Two items (simple):
ca = 2.0 $\longrightarrow$ num > 50% diameter narrowing, ψ = 0.1, α = 0.86, λ = 2.38
$exang$ = 1 $\longrightarrow$ num > 50% diameter narrowing, ψ = 0.21, α = 0.7, λ = 1.94
$thalach$ = 132.0 $\longrightarrow$ num > 50% diameter narrowing, ψ = 0.01, α = 1.0, λ = 2.75

component of the total cost [48]. According to our experience in designing and analysis of network decomposable algorithms, every algorithm step must have a number of exchanged messages for computing the various quantitative values. In this work and previous works [14]–[17], we use complexity models that involves the number of exchanged messages and reflecting the efficiency of decomposition carried out by the network algorithm. Due to the size of the messages, in this paper we consider Exchanging One Summary Per Message Model where only one local computation request is exchanged per message at a time.

In this section, for complexity analysis, we specify below an expression for the number of messages that need to be exchanged among the cooperated sites to generate the final association rules. Let us say there are n datasets, $D_1, D_2, \ldots D_n$, residing at n different network sites.

As seen above, frequent itemsets and association rules at each level of the algorithm can be determined by each site., i.e., the number of messages exchanged will be only n messages among the participating nodes. Then, the integrated association rules will be at coordinator site with no exchanged messages.

Finally for the validation phase, we need to validate the integrated association rules by each site, i.e., we need n exchanged messages for validation phase. Therefore, the total number of exchanged messages will be $2 * n$, i.e. we need to exchange a total of $2 * n$ messages among the stationary agents to run the association rule algorithm. This number of messages is not dependent on the number of tuples contained in each database and the system, therefore, is easily scalable to large databases. Also, this number of messages is much smaller than the data that may need to be transferred if we were to accumulate all databases at one site and then perform the data mining task. We have demonstrated above that the association rules algorithm can return the same results for distributed databases without having to move the databases to a centralized site. From the point of view of data security and privacy, the following observations can be made: No data tuple is exchanged between the sites and only the local association rules will be moved to the coordinator site.

E. SECURITY AND PRIVACY DISCUSSION

In the above, we proposed an algorithm for finding the association rules in a distributed database environment. In such environment, one would be interested in preserving the security and the privacy of the data. Our goal is to protect against eavesdroppers on the communication between the sites. We also would like to preserve the privacy of each site as much as possible.

1) EAVESDROPPING

Malicious attacker can listen to the communication between two different sites to obtain copy of the sent information. In this section, we show that anonymous listener cannot compromise the privacy of the transmitted information. To calculate the global results of support and confidence for the association rules, sites exchange information such as support and confidence, and number of tuples on site D_j that have the same attributes. To preserve the confidentiality of the information the proposed algorithm uses secure hash function to avoid sending the local results of site D_j. Each site will send support, confidence, and number of tuples of its local data as a hash digest. Each site D_j calculates the hash digest for its values as follow:

$$Hash\ Digest = h_A(Ks_{value}), \tag{12}$$

$$Hash\ Digest = h_A(Kc_{value}), \tag{13}$$

$$Hash\ Digest = h_A(Kn_{value}), \tag{14}$$

where K is a secret key that is shared and used by all the sites and s_{value}, c_{value}, n_{value} are the support and confidence, and number of tuples of D_j, respectively. This way the attacker or the listener cannot figure out the transmitted values. At the same time, since all sites are using the same key K and hash function $h_A()$ the coordinator site can still match hash digest values to form the global results. Anonymous listener cannot also figure out whether a specific site has a tuple with specific attribute value or not. The use of secret key in the hash makes it impossible for the anonymous listener to know the actual values that are sent.

Also, participating sites send counts of the transactions that contain the received item sets. Sites do not send actual values

of support, confidence, and number of tuples, but rather they multiply them by a secret constant. Thus, anonymous attacker who listens to the channel cannot figure out the original values of site j. Note that multiplying the support, confidence, and number of tuples values by secret constant will not affect the calculation of the global values because this secret value will be factored out at the coordinator site. Thus in summary, anonymous listener will not be able to figure out the the distinct values of the support, confidence, and number of tuples at any participating site.

2) SITE PRIVACY

This section discuss how much privacy each site can preserve even though the different sites need to collaborate to answer quires. More specifically we will address the the support, confidence, and number of tuples for each site.

Recall that all sites are using the same secret key. Site i cannot-figure out the support, confidence, and the number of tuples values of site j because the secure hash is a one way function. The participating sites send the hash of these values to coordinator site rather than sending the values themselves. If site i does not have a copy of the database schema, it cannot figure out the values that other sites send to the coordinator site.

In summary, the privacy of patients data is preserved since raw medical records are not exchanged between the local sites and the coordinator site. Only a secure summaries are exchanged.

F. EVALUATING THE PERFORMANCE OF THE PROPOSED ALGORITHM

In order to show the advantages of our Distributed Association Rule Mining for Predicting Heart Diseases algorithm (DRAM-PHD), we have performed a number of tests that demonstrate the ability of the proposed algorithm to work correctly in a distributed knowledge environment without moving all the databases to a single site. The tests were performed to find out the effect of various parameters on the final result. The two important variables that affect the result are:

1) the number of sites.
2) the number of tuples per database.

These tests have been carried out on a network of workstations connected by a LAN and tested against a number of databases of different sizes. We compare our results with the results of the algorithm in [7] where the authors solved the same problem in vertically distributed databases which will be same as horizontally distributed if all attributes are shared among the sites, i.e., all attributes are the same in all local databases. The algorithm in [7] has 2 versions: Optimized (Optimized PDMAR), where the coordinator site will simultaneously receive all results from the participating sites, while in the Un-optimized version (Un-optimized PDMAR), the coordinator site will receive the results from participating sites one at a time.

1) **Changing Number of Sites:** The first test was executed to demonstrate how the elapsed time and the number of exchanged messages varies with the number of local sites. In this test, we used the same setting as in [7], where number of participating sites varies between 2 and 20, and the time of sending a message is 0.125 second.

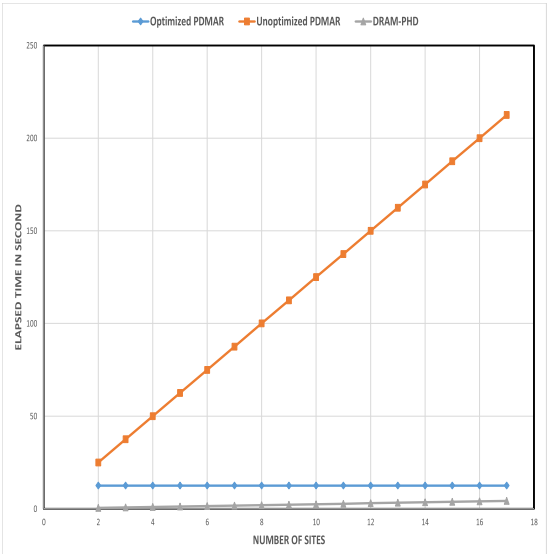

FIGURE 3. Message exchanges versus number of local sites in Optimized PDMAR, Un-optimized PDMAR and the proposed algorithm.

Figure 3 shows the number of exchanged messages between the participating sites and the coordinator site in the proposed algorithm, Optimized PDMAR and Un-optimized PDMAR. It can be seen easily that the number of exchanged messages increases as the number of local sites increases and the proposed algorithm has the smallest number of exchanged messages. This because the Optimized PDMAR and the Un-optimized PDMAR algorithms create a relation called shared and deal with each each set of tuples that correspondence to a shared tuple as a class. Creating shared relation and dealing with each class need more exchanged messages and so more elapsed time, while the proposed algorithm deals with all tuples at a participating site as a class. Figure 4 shows how the elapsed time to find the association rules in an implicit database D changes with the number of local sites in the proposed, Optimized PDMAR and Un-optimized PDMAR algorithms. It can be seen easily that the elapsed time increases as the number of local sites increases. and the proposed algorithm has the least elapsed time because the elapsed time depends on the number of exchanged messages.

2) **Changing the Total Number of Tuples in the Implicit Database:** In the second test, we demonstrate how the number of exchanged messages and the elapsed time vary with the number of tuples in the implicit database, i.e., with varying number of tuples at participating sites. Figure 5 shows the number of exchanged messages between the participating sites and the coordinator to find the association rules with varying the total num-

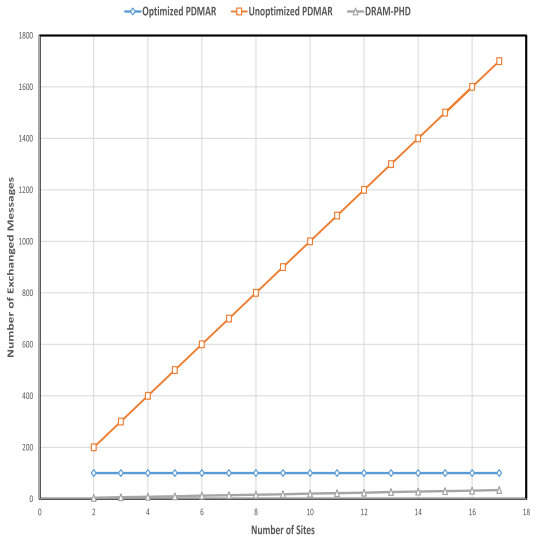

FIGURE 4. Elapsed time versus number of local sites in Optimized PDMAR, Un-optimized PDMAR and the proposed algorithm.

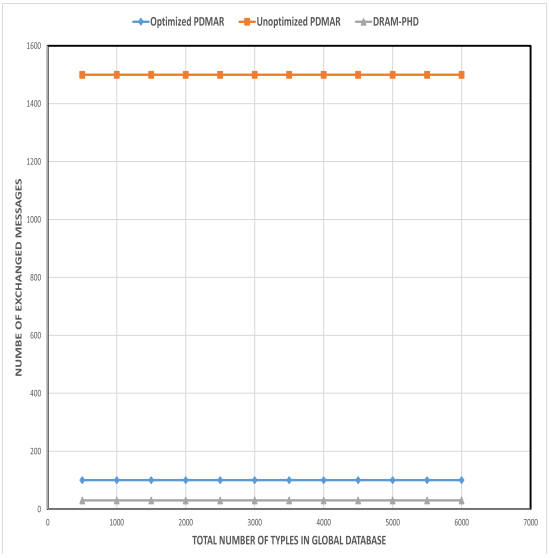

FIGURE 5. Exchanged messages versus number of tuples in the implicit database in Optimized PDMAR, Un-optimized PDMAR and the proposed algorithm.

ber of tuples in the implicit database D in proposed, Optimized PDMAR, and Un-optimized PDMAR algorithms. It can be seen easily that the number of exchanged messages increases as the number of local sites increases and the proposed algorithm has the least number of exchanged messages for same reasons above.

Figure 6 shows how the elapsed time to find the association rules in an implicit database D changes with the number of tuples at participating sites in proposed, Optimized PDMAR and Un-optimized PDMAR algorithms. It can be seen easily that the elapsed time increases as the number of participating sites increases and the proposed algorithm has the least elapsed time for the same reasons above.

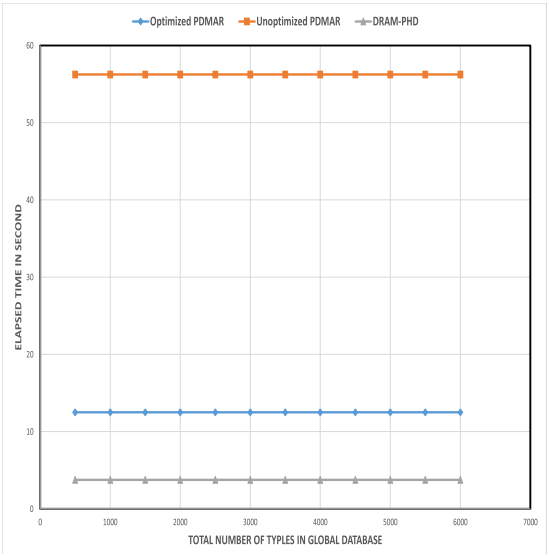

FIGURE 6. Elapsed time versus number of tuples in the implicit database in Optimized PDMAR, Un-optimized PDMAR and the proposed algorithm.

3) **Comparing with Traditional Algorithms:** In this test, we have compared the results of proposed algorithm with the results of the sequential algorithm in [3], which also discovers heart disease through association rules, in terms of the produced association rules and the elapsed time to produce these rules. The system in [3] has its dataset in one central site, while our proposed algorithm works on a dataset that is distributed among three participating sites. The algorithm succeeds in obtaining the same results to those that achieved by moving all the databases to one site, joining them, and then executing the traditional association rules algorithm as in [3].

From the above results, it is evident that the performance of our proposed algorithm continues to achieve the best results.

VI. CONCLUSION

Healthcare around the world is committed to providing quality care to patients via electronic health records. Due to the distributed nature of the EHRs, shared access to health records should be made possible and data integration should be established. Preserving the privacy of patient information is an important consideration when handling medical data. We have developed a privacy-preserving integration model based on association rules for predicting heart disease using patient data collected from horizontally distributed databases. Our model allows the sharing of data summaries (useful information) to be used to predict heart disease. These summaries are not accompanied by private patient information. Our approach is the first to use association rules metrics for naturally distributed medical datasets to generate weighted rules, which are further generalized using independent test datasets rather than using specific rules for each local model. In the future, work to discover more privacy-preserving integration models for vertically and horizontally distributed systems can be considered.

REFERENCES

[1] M. Eichelberg, T. Aden, J. Riesmeier, A. Dogac, and G. B. Laleci, "A survey and analysis of electronic healthcare record standards," *ACM Comput. Surveys*, vol. 37, no. 4, pp. 277–315, Dec. 2005.

[2] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: Towards better research applications and clinical care," *Nature Rev. Genet.*, vol. 13, no. 6, pp. 395–405, Jun. 2012.

[3] C. Ordonez, "Association rule discovery with the train and test approach for heart disease prediction," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 2, pp. 334–343, Apr. 2006.

[4] C. Ordonez, C. Santana, and L. Braal, "Discovering interesting association rules in medical data," in *Proc. ACM Data Mining Knowl. Discovery Workshop*, 2000, pp. 78–85.

[5] Y. Li, C. Bai, and C. K. Reddy, "A distributed ensemble approach for mining healthcare data under privacy constraints," *Inf. Sci.*, vol. 330, pp. 245–259, Feb. 2016.

[6] D. Tomar and S. Agarwal, "A survey on data mining approaches for healthcare," *Int. J. Bio-Sci. Bio-Technol.*, vol. 5, no. 5, pp. 241–266, Oct. 2013.

[7] A. Khedr, Z. A. L. Aghbari, and I. Kamel, "Privacy preserving decomposable mining association rules on distributed data," *Int. J. Eng. Technol.*, vol. 7, nos. 3–13, pp. 157–162, 2018.

[8] H. Jiawei, K. Micheline, and P. Jian, *The Morgan Kaufmann Series in Data Management Systems, Data Mining*, 3rd ed. San Mateo, CA, USA: Morgan Kaufmann, 2012, pp. 1–38.

[9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 1st ed. New York, NY, USA: Springer-Verlag, 2001.

[10] P. Clark and T. Niblett, "The CN2 induction algorithm," *Mach. Learn.*, vol. 3, no. 4, pp. 261–283, Mar. 1989.

[11] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[12] L. V. S. Lakshmanan, R. Ng, J. Han, and A. Pang, "Optimization of constrained frequent set queries with 2-variable constraints," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 1999, pp. 157–168.

[13] H. Yu, X. Jiang, and J. Vaidya, "Privacy-preserving SVM using nonlinear kernels on horizontally partitioned data," in *Proc. ACM Symp. Appl. Comput. (SAC)*, 2006, pp. 603–610.

[14] A. M. Khedr and R. Bhatnagar, "New algorithm for clustering distributed data using K-means," *Comput. Informat.*, vol. 33, pp. 1001–1022, Oct. 2014.

[15] A. M. Khedr, "Decomposable naive Bayes classifier for partitioned data," *Comput. Informat.*, vol. 31, pp. 1511–1531, Jan. 2012.

[16] A. M. Khedr, "Nearest neighbor clustering over partitioned data," *Comput. Informat.*, vol. 30, pp. 1001–1026, Jan. 2011.

[17] A. M. Khedr, "Learning K-classifier from distributed databases," *Comput. Informat. J.*, vol. 27, no. 3, pp. 355–376, 2008.

[18] R. Au and P. Croll, "Consumer-centric and privacy-preserving identity management for distributed E-health systems," in *Proc. 41st Annu. Hawaii Int. Conf. Syst. Sci. (HICSS)*, Jan. 2008, pp. 1–10.

[19] O. R. L. Sheng and H.-M.-C. Garcia, "Information management in hospitals: An integrating approach," in *Proc. 9th Annu. Int. Phoenix Conf. Comput. Commun. Conf.*, 1990, pp. 296–297.

[20] A. Jha, Z. Li, E. Orav, and A. Epstein, "Care in US hospitals—The hospital quality alliance program," *New England J. Med.*, vol. 353, pp. 265–274, Jul. 2005.

[21] C. Ahn, Y. Nah, S. Park, and J. Kim, "An integrated medical information system using XML," in *The Human Society and the Internet Internet-Related Socio-Economic Issues* (Lecture Notes in Computer Science), vol. 2105, W. Kim, T. W. Ling, Y. J. Lee, and S. S. Park, Eds. Berlin, Germany: Springer, 2001.

[22] S. Ajami and T. Bagheri-Tadi, "Barriers for adopting electronic health records (EHRs) by physicians," *Acta Inf. Med.*, vol. 21, no. 2, pp. 129–134, 2013.

[23] American Health Information Management Association, "The 10 security," 2013.

[24] *US Department of Health and Human Services. Security Standards: General Rules*, document 46 CFR Section 164.308(a)-(c).

[25] D. Meinert, "Resistance to electronic medical records (EMRs): A barrier to improved quality of care," *Informing Sci., Int. J. Emerg. Transdiscipline*, vol. 2, pp. 493–504, Jan. 2004.

[26] E. W. Ford, N. Menachemi, and M. T. Phillips, "Predicting the adoption of electronic health records by physicians: When will health care be paperless?" *J. Amer. Med. Inform. Assoc.*, vol. 13, no. 1, pp. 106–112, Jan. 2006.

[27] M. A. Cifci and S. Hussain, "Data mining usage and applications in health services," *Int. J. Informat. Vis.*, vol. 2, no. 4, p. 225, Jun. 2018, doi: 10.30630/joiv.2.4.148.

[28] R. Ray, "Advances in data mining: Healthcare applications," *Int. Res. J. Eng. Technol.*, vol. 5, no. 3, pp. 2356–2395, Mar-2018.

[29] S. Gambs, B. Kégl, and E. Aïmeur, "Privacy-preserving boosting," *Data Mining and Knowledge Discovery*, vol. 14, no. 1, pp. 131–170, 2007.

[30] P. Nayak, "A survey on medical data by using data mining techniques," *Int. J. Advance Res., Ideas Innov. Technol.*, vol. 3, no. 6, pp. 1330–1335, 2017.

[31] D. Roffman, G. Hart, M. Girardi, C. J. Ko, and J. Deng, "Predicting non-melanoma skin cancer via a multi-parameterized artificial neural network," *Sci. Rep.*, vol. 8, no. 1, p. 1701, Dec. 2018.

[32] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017.

[33] R. Subhashini and M. K. Jeyakumar, "OF-KNN technique: An approach for chronic kidney disease prediction," *Int. J. Pure Appl. Math.*, vol. 116, no. 24, pp. 331–348, 2017.

[34] N. Nahar and F. Ara, "Liver disease prediction by using different decision tree techniques," *Int. J. Data Mining Knowl. Manage. Process*, vol. 8, no. 2, pp. 1–9, Mar. 2018.

[35] K. Mathan, P. M. Kumar, P. Panchatcharam, G. Manogaran, and R. Varadharajan, "A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease," *Des. Automat. Embedded Syst.*, vol. 22, no. 3, p. 225, 2018.

[36] C.-H. Hsu, G. Manogaran, P. Panchatcharam, and S. Vivekanandan, "A new approach for prediction of lung carcinoma using back propagation neural network with decision tree classifiers," in *Proc. IEEE 8th Int. Symp. Cloud Service Comput. (SC)*, Nov. 2018, pp. 111–115.

[37] E. Avci, "A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier," *Expert Syst. Appl.*, vol. 36, no. 7, pp. 10618–10626, Sep. 2009.

[38] I. Vidić, L. Egnell, N. P. Jerome, J. R. Teruel, T. E. Sjøbakk, A. Østlie, H. E. Fjøsne, T. F. Bathen, and P. E. Goa, "Support vector machine for breast cancer classification using diffusion-weighted MRI histogram features: Preliminary study: Machine learning in DWI of breast cancer," *J. Magn. Reson. Imag.*, vol. 47, no. 5, pp. 1205–1216, May 2018.

[39] C.-L. Huang, H.-C. Liao, and M.-C. Chen, "Prediction model building and feature selection with support vector machines in breast cancer diagnosis," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 578–587, Jan. 2008.

[40] S. Ramasamy and K. Nirmala, "Disease prediction in data mining using association rule mining and keyword based clustering algorithms," *Int. J. Comput. Appl.*, vol. 42, no. 1, pp. 1–8, Jan. 2020.

[41] R. Chaves, J. M. Górriz, J. Ramírez, I. A. Illán, D. Salas-Gonzalez, and M. Gómez-Río, "Efficient mining of association rules for the early diagnosis of Alzheimer's disease," *Phys. Med. Biol.*, vol. 56, no. 18, pp. 6047–6063, Sep. 2011.

[42] A. A. Bakar, Z. Kefli, S. Abdullah, and M. Sahani, "Predictive models for dengue outbreak using multiple rulebase classifiers," in *Proc. Int. Conf. Electr. Eng. Informat.*, Jul. 2011, pp. 1–6.

[43] A. Anguita, D. Pérez-Rey, J. Crespo, and V. Maojo, "Automatic generation of integration and preprocessing ontologies for biomedical sources in a distributed scenario," in *Proc. 21st IEEE Int. Symp. Comput.-Based Med. Syst.*, Jyväskylä, Finland, Jun. 2008, pp. 336–341.

[44] A. Khedr and R. Bhatnagar, "Agents for integrating distributed data for complex computations," *Comput. Informat. J.*, vol. 26, no. 2, pp. 149–170, 2007.

[45] M. Kantarcoglu, J. Vaidya, and C. Clifton, "Privacy preserving naive Bayes classifier for horizontally partitioned data," in *Proc. IEEE ICDM Workshop Privacy Preserving Data Mining*, Nov. 2003, pp. 3–9.

[46] R. Sujatha and A. Nithya, "A survey of health care prediction using data mining," *Int. J. Innov. Res. Sci., Eng. Technol.*, vol. 5, no. 8, p. 14538, 2016.

[47] W. A. AlZoubi, "Mining medical databases using graph based association rules," *Int. J. Mach. Learn. Comput.*, vol. 3, pp. 294–296, Jun. 2013.

[48] C. Wang and M.-S. Chen, "On the complexity of distributed query optimization," *IEEE Trans. Knowl. Data Eng.*, vol. 8, no. 4, pp. 650–662, Aug. 1996.

[49] L. I. Chengyan, S. Feng, and G. Sun, "DCE-miner: An association rule mining algorithm for multimedia based on the MapReduce framework," *Multimedia Tools Appl.*, vol. 79, pp. 1–23, Jun. 2020.

[50] P. Y. Taşer, K. U. Birant, and D. Birant, "Multitask-based association rule mining," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 28, no. 2, pp. 933–955, Mar. 2020.