IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.
Digital Object Identifier 10.1109/ACCESS. 2017.Doi Number

Privacy Preserving Data Mining Framework for Negative Association Rules: An Application to Healthcare Informatics

SAAD M. DARWISH[1], REHAM M. ESSA[2], MOHAMED A. OSMAN[2], AND AHMED A. ISMAIL[3]
[1] Department of Information Technology, Institute of Graduate Studies and Research, University of Alexandria, 21526, Egypt
[2]Management Information System Department, Higher Institute of Management Information Technology, Kafr el-Sheikh, 33511, Egypt
[3] The Higher Institute of Computers and Information Systems, Abo Qir, Alexandria 21913, Egypt.

Corresponding author: Saad M. Darwish (e-mail: saad.darwish@alexu.edu.eg).

ABSTRACT Protecting the privacy of healthcare information is an important part of encouraging data custodians to give accurate records so that mining may proceed with confidence. The application of association rule mining in healthcare data has been widespread to this point in time. Most applications focus on positive association rules, ignoring the negative consequences of particular diagnostic techniques. When it comes to bridging divergent diseases and drugs, negative association rules may give more helpful information than positive ones. This is especially true when it comes to physicians and social organizations (e.g., a certain symptom will not arise when certain symptoms exist). Data mining in healthcare must be done in a way that protects the identity of patients, especially when dealing with sensitive information. However, revealing this information puts it at risk of attack. Healthcare data privacy protection has lately been addressed by technologies that disrupt data (data sanitization) and reconstruct aggregate distributions in the interest of doing research in data mining. In this study, metaheuristic-based data sanitization for healthcare data mining is investigated in order to keep patient privacy protected. It is hoped that by using the Tabu-genetic algorithm as an optimization tool, the suggested technique chooses item sets to be sanitized (modified) from transactions that satisfy sensitive negative criteria with the goal of minimizing changes to the original database. Experiments with benchmark healthcare datasets show that the suggested privacy preserving data mining (PPDM) method outperforms existing algorithms in terms of Hiding Failure (HF), Artificial Rule Generation (AR), and Lost Rules (LR).

INDEX TERMS Privacy-preserving data mining, healthcare data, evolutionary computation, sanitization process.

I. INTRODUCTION

Electronic health records (EHRs) are widely used by a variety of healthcare organizations in an effort to enhance patient care and enhance the efficiency of healthcare delivery. In complex clinical environments, the EHR system accelerates the clinician's workflow by automating the data management process. When utilized effectively, these EHRs not only facilitate many routine health care tasks, but also help in the accurate identification of diseases. Individuals' access to their medical records is facilitated by EHRs. In addition, they come with a home health monitoring system that allows patients to measure and evaluate their symptoms every day [1].

The dissemination of data from the EMR system is critical for improving the quality of medical research. Researchers use this data to perform a wide range of tasks involving data mining [2], such as classification (prediction of diabetic presence) [3], clustering (risk identification) [4], statistical tests (body mass index and diabetes association) [5], or query responding. In addition to enhancing the actualized human services to patients, researchers in healthcare are expected to benefit from the integration of information and electronic

health records (EHRs). Fig.1 illustrates the main data mining applications in healthcare; see [3] [4] for more details.

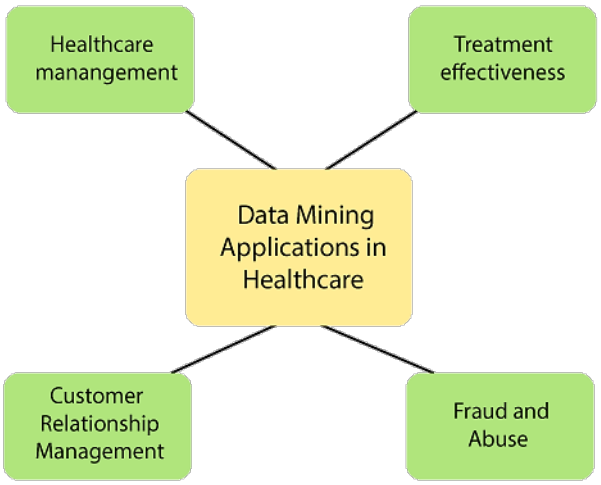

FIGURE 1: **Data mining applications in healthcare.**

Anonymization-based methods and cryptographic methods have both been presented in the literature as means of achieving privacy [1] [3]. Anonymization is widely used by researchers owing to the lower communication and computation costs of the same as compared to their cryptographic counterparts. Information loss is one of the important issues in an anonymization-based approach. The information loss calculates the difference between the original databases and the anonymized databases. The information loss increases with the increase in the level of the generalization and/or suppression method. Generally, the information loss should be less to achieve higher data utility [6-8].

K-anonymity is a key concept that was introduced to address the risk of re-identification of anonymized data through linkage to other datasets. K-Anonymity is able to prevent identity disclosure, i.e., a record in the k-anonymized data set cannot be mapped back to the corresponding record in the original data set. However, in general, it may fail to protect against attribute disclosure. Some critics of k-anonymization take issue with the fact that achieving a re-identification risk of zero is impractical or impossible [8-11].

The vast majority of approaches fall into two broad categories: those that protect sensitive data during mining and those that protect sensitive data mining outputs. The first category includes methods like perturbation, sampling, and modification to generate sanitized datasets that may be safely shared with other parties. These strategies are designed to help data miners get valid results even when they don't have any real data. Contrarily, the second category includes methods for keeping sensitive knowledge patterns obtained from the use of data mining algorithms secret, including methods for lowering the performance of classifiers in classification tasks so that sensitive information is not revealed [1][3][8].

With the advent of significant data mining tools and systems, we are faced with the challenge of building a healthcare data mining system that meets user expectations for meaningful knowledge discovery from databases while avoiding the potential to infer personal information about people. A person or organization should not be able to be identified (by a third party) based on the mining techniques or results that we share [1] [2]. Preventing sensitive information from being released in an unauthorized or unsolicited way is the goal of Privacy-Preserving Data Mining (PPDM). In healthcare systems, the PPDM method is very beneficial for minimizing the leakage of sensitive personal information. Thus, it enables the analysis of enormous data sets from medical research systems in order to obtain information and identify therapeutic methods for deadly illnesses without invading patients' privacy. Although database management systems have always prioritized data security, mining knowledge and restricting the exposure of sensitive information has now become the most crucial and highest priority objective of the data mining process [3][4][9].

Privacy-preserving approaches try to avoid the leakage of sensitive data, making it more difficult to deduce sensitive data from non-sensitive data [9]. However, they do not exclude the discovery of inference rules. As a result, experts have focused their attention on privacy-preserving association rules in recent years. Association rule mechanisms have been extensively employed in a large variety of enterprises and industrial firms across a wide variety of industry sectors, including marketing, forecasting, diagnostics, and security [10][11]. Sensitive association rule concealment is a subsection of PPDM that falls under the category of output privacy. Restrictive rules are those that should be kept secret. One of the PPDM techniques used to keep private information hidden is known as data sanitization.

Database sanitization algorithms that protect sensitive non-production database data from unauthorized access are being developed and tested using a range of approaches currently available in the literature [9]. These algorithms fall under the following categories: (1) Algorithms control the concealing process using the rule's support or confidence; and (2) Algorithms change raw data by distorting or blocking the original values. It is possible that the modification process alters the original set of rules that may be mined from the original database, either by hiding non-sensitive rules (lost rules) or by introducing rules that were not supported by the original database into the mining process (ghost rules). We attempted to mitigate these unfavorable effects by modifying the original dataset a little and appropriately. Because of the difficulty of PPDM, which is called an NP-hard problem, it's hard to find the best solution with the fewest side effects.

Using negative association criteria to find overlapping symptoms or drugs that work well together is a useful tool in healthcare data analysis. Negative association rule mining is difficult to accomplish because of the underlying differences between positive and negative association rule mining. When

it comes to mining negative association rules, there are two primary concerns that need to be handled by the researchers: looking for and filtering the negative association rules [9-11].

A Genetic Algorithm (GA) looks for a solution using a population of points rather than a single point. This approach is both computationally straightforward and effective. Tabu Search (TS) examines each string as a point in the solution space. TS leads iterations from one neighborhood point to another by enhancing the solutions locally and has the ability to avoid starved local minima. Combining GA and TS with their respective strengths increases the likelihood of finding a suitable solution to global combinatorial optimization issues. GA begins with a collection of preliminary solutions and generates a set of innovative solutions using the hybrid search approach. TS develops each set of unique solutions via a local search. GA then uses the augmented solution of TS in order to keep the same evolution [12].

A. PROBLEM STATEMENT

The researchers noted that the integrated health care system has evolved into a critical component of the current health information system. Medical personnel may gain from data mining since it allows them to expand their practicability by sharing and evaluating results with others. This type of information demands a higher level of privacy, which ensures that the association rules remain safe even when data owners use a shared cloud. PPDM has become a significant concern in recent years due to its ability to conceal not just private information but also enable the discovery of essential information using various data mining methods. PPDM is a so-called NP-hard issue. The reason for this is that traditional PPDM algorithms are primarily concerned with concealing sensitive information to the greatest extent feasible. This development might have significant unexpected consequences in terms of missing and artificial costs. Since both side effects are taken into account, it is difficult to choose the best technique.

B. MOTIVATION

When using healthcare data, it is essential to weigh the benefits of data privacy against the potential downsides. For the positive association rule in PPDM data sanitization, several heuristics and metaheuristics-based approaches have been developed in the past. Additions and deletions are used to hide critical information in the original database. An NP-hard problem of this kind is well-known. PPDM issues have been solved using heuristic methods including greedy search, meta-heuristic approaches like the Genetic Algorithm (GA), and Particle Swarm Optimization (PSO). Most of these traditional methods are static and need a significant amount of processing.

Negative association rules mining is not a simple issue to solve, since it is widely known that a significant number of uninteresting rules are created, and it is necessary to do extra work in order to choose just the best rules. Applying a GA to optimize extracted negative rules incurs additional time and space costs due to the random selection of the starting population of rules. A GA's fundamental shortcoming is its unguided mutation. In GA, the mutation operator acts similarly to adding a randomly generated number to a parameter of a population member. This is the only reason for the GAs' much delayed convergence. Integration with other algorithms that guide search and metaheuristic-based data sanitization may be able to solve this problem.

C. CONTRIBUTION

Despite their enormous success, meta-heuristics optimization tools have never been employed to address the PPDM problem associated with negative association rules in healthcare data. In this work, the issue of negative association rules in healthcare data mining is addressed using a novel approach based on the Tabu-Genetic (GTA) optimization framework. The suggested architecture is flexible, and it hides sensitive data by deleting it. This makes sure that personal information is safe, but also allows for new information to be found.

Unlike prior approaches in this area, this novel method develops so-called “meta-heuristic” negative association rules using a combination of genetic algorithms and Tabu search to increase the perfection of these rules in a time and memory-efficient manner. The advantage of Tabu search is that it eliminates a large number of redundant rules and item sets. It is possible to dynamically hide sensitive information using a deletion operation-based perturbation approach, rather than pre-defining it. The algorithm's key contributions are as follows:

- This is the first work to handle the PPDM problem for negative association rules using a Tabu-Genetic technique that outperforms a traditional approach in terms of rule hiding side effects.
- As a way to speed up the evolution process, the Tabu concept is used to speed up the evaluation of the solution that has been tested. This reduces the need for extra database scans.
- Rather than predefining the transactions to be disrupted for information concealment, the employed technique dynamically identifies them.

The rest of the article is organized in the following manner. The second section highlights related work. Section III describes the proposed method for sanitizing healthcare data mining. The experimental results are presented in Section IV. Section V contains a conclusion and a discussion on future work.

II. RELATED WORK

Healthcare process data can include a large number of sensitive variables and highly changeable process behaviors that pose extra privacy issues. The healthcare industry must conform to strict data privacy standards. Privacy protection for

such data while maintaining its usefulness for process mining is an ongoing concern in healthcare. There is a trade-off between data privacy and usefulness in the use of existing data transformation methods for anonymizing healthcare process data. For example, encryption does not provide enough privacy protection when used to optimize the value of data for process mining. The accuracy of results may be compromised if methods that conform to more severe privacy rules (such as generalization) are used [1-3] [8].

In literature, major research has used anonymity, data masking, data perturbation, and cryptography for data privacy. Using dynamic data masking, we are able to achieve format-preserving masking and anonymization without having to manually copy data or remove values—tasks which can not only delay analysis, but can weaken the utility of data and introduce the risk of human error. The cryptographic approach is especially difficult to scale when more than a few parties are involved. It also does not address the question of whether disclosing the final data mining results may violate the privacy of individual records. The perturbation approach does not reconstruct the original data values. New algorithms have been developed to reconstruct the original data distribution. In general, every technique has its own demerits, i.e., information loss, privacy breach, and low data utility [3] [6].

It is easy to see that anonymity is not enough. For example, suppose we use *k*-anonymity to protect data. This means that, knowing identifying information about an individual, there are at least *k* records in the database that could (with equal probability) refer to that individual. However, suppose that those records also include sensitive information, e.g., if an individual is diabetic. If all *k* individuals have the same value for sensitive information (for example, all are diabetic), then *k*-anonymity offers no protection against disclosure of that fact. This has led to alternate approaches. However, it is still difficult to answer the question, "is the data anonymous enough?" [3].

A number of sophistications of *k*-anonymity have been suggested, like *p*-sensitive *k*-anonymity, *l*-diversity, and *t*-closeness. A data set is said to satisfy *l*-diversity if there are at least *l* well-represented values for each confidential attribute in each group of records that share key attributes. However, this extension suffers from a similarity attack. If the values of a sensitive attribute in a group are l-diverse but the semantics are the same, the attribute will also be revealed [6] [7]. But the work in [8] solves this problem by proposing (l, d)-semantic diversity, which extends l-diversity. See [8] for more details.

Regarding, *p*- sensitive *k*-anonymity, its purpose is to protect against attribute disclosure by requiring that there be at least *p* different values for each confidential attribute within the records sharing a combination of key attributes. *P*-sensitive *k*-anonymity has the limitation of implicitly assuming that each confidential attribute takes values uniformly over its domain, that is, that the frequencies of the various values of a confidential attribute are similar. When this is not the case, achieving *p*-sensitive *k*-anonymity may cause a huge data utility loss. *t*-closeness solves the attribute disclosure vulnerabilities inherent to *l*-diversity: Skewness attack, since the within-group distribution of confidential attributes is the same as the distribution of those attributes for the entire dataset, no skewness attack can occur. Similarity attack, again, since the within-group distribution of confidential attributes mimics the distribution of those attributes over the entire dataset, no semantic similarity can occur within a group that does not occur in the entire dataset. Of course, within-group similarity cannot be avoided if all patients in a data set have similar diseases [3][6][8].

Data mining techniques for privacy protection can be general or specific [13] [14]. Data mining operations may employ generic approaches to transform data into a form that can be used as an input. Changing records without adding new values or changing existing values may be utilized to achieve anonymity using these methods (e.g., data swapping) (e.g., by adding noise). Certain data mining techniques include privacy protections in their algorithms (e.g., privacy-preserving decision tree classification). Anonymization methods (such as association rule hiding) have been described for sensitive data mining outputs as well [4].

A machine learning approach is used to link heterogeneous data based on privacy-preserving data mining [14]. In [15], the authors discussed the use of hierarchical categorization approaches in PPDM. Dasseni et al. [16] established a Hamming distance-based strategy to diminish the support or confidence in sensitive information. Oliveira and Zaane [17] created many sanitization techniques that were used to conceal frequently occurring itemsets using a heuristic approach. This method was utilized to avoid the inclusion of noise and minimize the quantity of genuine data that was deleted from the dataset. Islam and Brankovic [18] suggested an approach for protecting and concealing individual privacy while retaining excellent data quality using the noise addition approach. In [19], the authors described a sanitization technique in which the victim item is the item that is used the most often in the transaction. In addition, a threshold for sharing is set up to find a balance between privacy and information sharing.

Sun and Philip [20] suggested using the boundary of non-sensitive itemsets to monitor the perturbation process's influence on the sanitized database. In [21], the authors used the MaxMin method to accomplish the same thing. In this situation, the sanitized database's quality is maintained by picking the least impactful transaction for change at each stage [21]. For the purpose of concealing sensitive data, Amiri [22] outlined three heuristic strategies: aggregate, disaggregate, and hybrid. The aggregate approach lowers the amount of support needed for a sensitive itemset by eliminating some transactions from the database. The disaggregated technique reduces the amount of support for sensitive items by removing specific items from the list of items. Two previous methods were combined to create the hybrid approach. The aggregate technique is used to determine which transactions are

victimized. Then, using the disaggregation method, the victim items from the transactions that were chosen are changed.

Wang et al. [23] established two approaches for hiding informative association rules. Rather than hiding critical association rules, a list of predicted items is presented. Then, the informative association rules whose antecedent contains the anticipated items cannot be mined from the resulting database. Wu et al. [24] developed a method for tracking all potentially significant changes. The database is then changed in a hidden way with the fewest possible side effects from the template until all sensitive rules are completely hidden. Gkoulalas et al. [25] presented an approach for maintaining sensitive itemsets based on the use of borders. By expanding the original database, it is possible to limit support for sensitive items.

Wu et al. [26] presented two greedy-based algorithms for the purpose of preserving sensitive association rules: the greedy approximation technique and the greedy exhausting approach. Both strategies obfuscate the critical rules by creating or removing certain database objects. The contrast is that the latter takes the database's effect into account when calculating expenses. Cheng et al. [27] discussed the advantages and disadvantages of positive and negative border rules. These ideas are used to determine the likelihood that a rule may become a missing or false rule after data sanitization. Each sensitive transaction is assigned two meaningful values based on the positive and negative border rules. To disguise the sensitive association rules, the least relevant transaction is modified. Hong et al. [28] used the term frequency-inverse document frequency approach for developing a strategy for reducing the support for sensitive item sets. A transaction containing a large number of sensitive items but having minimal influence on other transactions is very likely to be updated. The deletion priority is determined by the number of sensitive items supported.

Because the advancement of the PPDM is generally an NP-hard issue, it is preferable to provide meta-heuristic methodologies for determining the best solutions. The authors of [29] devised a secure mechanism for employing evolutionary algorithms to find a more optimal set of rules without disclosing their private data. The encoded chromosome is seen as a collection of solutions in this situation, and the transaction of a gene inside a chromosome is viewed as the victim of later deletion. Additionally, a fitness function was constructed to account for three side effects during an assessment using predefined weights to demonstrate the chromosome's quality. The same idea was introduced in [30], with different chromosome representations and different fitness functions. Although the algorithms stated above are effective at selecting the optimal transactions for deletion, pre-defined weights for side effects are still required; this step has the ability to dramatically modify the final outcomes of the suggested systems. They are efficient. As a result of the sanitization procedure, the lost and ghost rules will be produced at random once the item has been randomly updated.

A multi-objective optimization (e.g. NSGA II) approach was developed by some researchers to solve the above concerns, taking into consideration both data and knowledge distortion. However, this technique integrates multi-objective functions, yet it may result in inadequate information for decision-making since it directly deletes attributes from databases. The sequential dataset does not support this claim [15]. In contrast to the traditional particle swarm optimization method, multi-objective particle swarm optimization (MOPSO) makes use of a number of additional factors. MOPSO, on the other hand, cannot be used directly to address the PPDM problem because dominant relationships must be leveraged to obtain the best deletion transactions [30–32].

In [31], the authors used the MOPSO framework to present a hierarchical-cluster algorithm for concealing sensitive itemsets. Although partial transactions might be generated as a result, this can result in a misleading decision, especially in the handling of hospital diagnoses. Recently, a deep reinforcement learning technique has been applied to sanitize sensitive data from a database while still protecting privacy and allowing for knowledge discovery [33] [34]. For more information in this area, the reader can refer to the recent research in [35–39].

According to what is stated in [59-64], a new algorithm for quickly hiding sensitive association rules has been developed. Within this algorithm, a heuristic function is used to further determine the earlier weight for each specific transaction. This allows for the order of modified transactions to be decided in an efficient manner. Therefore, the links between the sensitive association rules and each transaction in the primary database are analyzed by successfully selecting the appropriate item for alteration. This is done so that any necessary changes may be made. Another strategy was offered, in which the sensitive association rules were concealed with the use of a novel multi-objective method, which ultimately resulted in an increase in the database's level of safety. This approach is derived on the idea of a genetic algorithm, which improves both the confidentiality of the dataset and its precision.

In addition, an accurate border-based method was used to achieve an ideal solution to conceal sensitive frequent item sets with a minimal extension of the original database generated synthetically. The development of the database extension is formulated by the system as a constraint satisfaction problem, and the mapping of constraint satisfaction concerns to an analogous binary integer programming problem is used. A new approach that is item-set oriented has been described. In this algorithm, the support for large item-sets is drastically decreased to a level that is much below the threshold that the client has specified. As a result, there is no way to derive any rules from the individual item sets. A novel method for selecting the items that need to be removed from the dataset in order to prevent the detection of a set of rules is also presented as part of this research. Most of the problems are caused by choosing victim-items without making changes to patterns that aren't sensitive. The notion of

representative association rules is used by several algorithms in order to locate potentially sensitive items. The heuristic for confidence and support reduction based on the intersection lattice method is used in order to cut down on the negative impacts that are caused by the process. By minimizing the modifications on database the efficiency can be enhanced with reduced side effects.

A. THE NEED TO EXTEND THE RELATED WORK

Numerous authors proved that the optimum sanitization issue is NP-hard and developed a heuristic technique for hiding sensitive frequently occurring item sets. The PPDM problem has led to the development of several evolutionary algorithms. Nonetheless, these algorithms rely heavily on data cleaning-based sanitization methods for positive association rules. When it comes to negative association rules, data sanitization presents numerous key challenges. To begin, typical healthcare transaction databases include hundreds of drugs, but only a handful of them are contained in each record (patient transaction). If a database contains 10,000 drugs and each patient is treated with an average of ten of them, the database's density is 10% of the total. From the standpoint of negative patterns (showing the absence of drugs), the density rises to 99.9 %, resulting in an explosion of rules, the majority of which are uninteresting. Second, the difficulty of association rule mining methods increases exponentially with the number of items; if a negated item is considered for each item in the database, the computation costs increase.

The approach proposed in this article attempts to address the constraints of existing sanitization algorithms by concealing particular itemsets (chosen using the Tabu-genetic method) rather than rules. The suggested method hides rules without adding to the original database (there are no fake transactions) and with the least number of lost or ghost rules.

III. PROPOSED SANITIZATION ALGORITHM FOR NEGATIVE RULES

By replacing anonymous values or omitting critical attribute values, the privacy of public healthcare data is often preserved [1]. Fig. 2 depicts the suggested technique for sanitizing healthcare databases. The suggested database sanitizing technique uses a genetic algorithm to choose the ideal items to modify in order to hide sensitive negative association rules, as opposed to previous efforts that employed negative association rules and hidden specific items rather than specific rules. However, hiding itemsets prevents them from appearing in any rules exceeding the minimum confidence level, regardless of whether those rules are sensitive or non-sensitive. However, hiding specific rules attempts to modify itemsets contained in these rules so that sensitive rules can be reduced to a user-specified threshold, thereby allowing the same items that appear in insensitive rules to appear in other non-sensitive rules, resulting in an increase in false positives. For problem formulation, see [12][14][15][24][26][29] for more details.

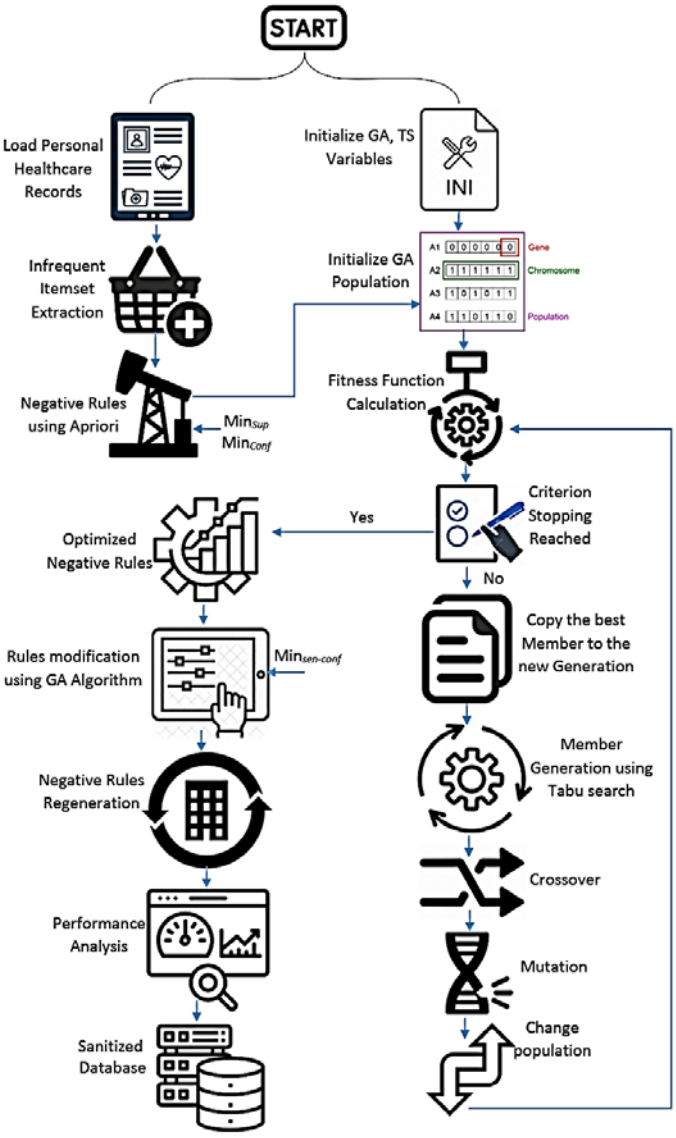

FIGURE 2: The proposed sanitization algorithm for negative association rules based on genetic-Tabu heuristic search.

Using a binary transactional dataset as an input, this work uses evolutionary algorithms to change it such that all sensitive rules are concealed and the original dataset is only slightly altered. Corrupting the original database (i.e., replacing 1s with 0s and vice versa) is the most common way of modifying transactions. This research builds on our prior work [12] by merging GA and Tabu Search to tackle a mining negative association rule issue (TS). The suggested technique increases the accuracy of mined results throughout the mining phases by using negative interestingness and negative confidence. Through genetic-Tabu search, pruning is used to eliminate uninteresting negative association rules. This strategy outperforms others by using efficient chromosomal representation and neighborhood tactics. We will detail our strategy in the next section. The most critical aspect of this effort is defining our fitness function in GA.

PHASE 1: EXTRACTING OPTIMIZED NEGATIVE RULES

- Convert an input dataset containing a collection of attributes (items) and instances (records) to numeric values (coding).
- For a variable-length item set, generate all infrequent item sets using the Apriori method; for more information, see [12] [40-43]
- From infrequent item sets, build initial negative association rules using the Apriori technique.
- Define the fitness function for specific interesting negative association rules.
- The GA is used to construct the chromosomes associated with the negative association rules and to determine the fitness value of each individual chromosome at this point. Build negative association rules based on how each chromosome relates to the average fitness value.
- For each hybrid search, GA generates new solutions based on the initial solutions it generates. In order to improve each new set of solutions, Tabu search does a local search. Then GA continues parallel development using the better solution of TS. Tabu Search examines each string as a point in the solution space. TS directs iterations from one neighborhood point to the next by enhancing the solution's quality locally and by avoiding suboptimal local minima. The combination of GA and TS, each of which has distinct advantages, offers a fair possibility of giving a viable solution to global combinatorial optimization issues.
- Reset the fitness value, compute the final negative rules, and modify the remaining child chromosomes after crossover and mutation. The following is a metric for discovering interesting negative association rules [8].

$$X \rightarrow \neg Y \ : \ \frac{N_b}{N_a + N_b} \tag{1}$$

$$\neg X \rightarrow Y \ : \ \frac{N_c}{N_c + N_d} \tag{2}$$

$$\neg X \rightarrow \neg Y \ : \ \frac{N_d}{N_c + N_d} \tag{3}$$

where N_a denotes the number of transactions including both X and Y, N_b is the number of transactions containing only X, N_c denotes the number of transactions containing just Y, and N_d specifies the number of transactions containing neither X ($\neg X$) nor Y ($\neg Y$). Rule discovery should have a high degree of prediction accuracy, be easy to understand, and be engaging for the reader [12] [42-45]. The fitness function should be tailored to the particular search areas; consequently, the fitness function used is critical for obtaining the required results. It is well established that the rule is more effective when N_a and N_d are greater than N_b and N_c [42][46].

$$Confidence\ Factor\ (\mathrm{CF}) = \{N_a / (N_a + N_c)\} \tag{4}$$

The suggested model also relies on another factor, completeness ($Comp$) measure for computing the fitness function.

$$Comp = \{N_a / (N_a + N_b)\} \tag{5}$$

$$\text{Fitness} = (\mathrm{CF} * Comp) \tag{6}$$

In this scenario, the system encodes the rules using the same form as in [46]. The best of the new solutions is then chosen for further testing using Tabu restrictions and aspiration criteria. If they pass the previous test, they may be added to the Tabu list memory. If they fail the former criteria but pass the latter, they may also be stored in the Tabu list memory. The crossover operator is used to exchange genetic material (bit-values) between the two parent strings given the optimal population obtained by the Tabu search stage (selection operator). It is conceivable that a crossover operation will result in the formation of a degenerate population. A mutation operation is used to reverse this. For further information, see [12].

As mentioned before, the proposed model combines modules from previous research [9] [12] [42] [47] [48]. Phase 1 in our work is the same as Phase 1 in our previous work [12] that uses Tabu-genetic to extract both positive and negative association rules. However, in our work, we focused on extracting negative rules only with the aim of reducing computational cost and highlighting the importance of these rules within the healthcare system. In [42], the authors extract positive and negative rules from the Apriori algorithm without utilizing an optimization procedure to extract only interesting rules. In our algorithm, we use an optimization method that combines Tabu search and genetic algorithm to find only interesting rules and cut down on the cost of computing.

PHASE 2: GENETIC ALGORITHM- BASED DATA SANITIZATION

This step is when the suggested framework for hiding sensitive negative association rules makes its most significant contribution. As a result of the preceding phase's sensitive rules, the system tries to hide them by decreasing their confidence to less than the predefined threshold by raising support for the antecedent and decreasing support for the consequent through the substitution of 1's for 0's in the transactions [16][27][40]. As a result of modifying all sensitive item sets related to sensitive rules in all database transactions, the algorithm will need a lot of processing power. In our present study, we address the aforementioned issue by using GA to identify the optimum itemsets for alteration. As a result, there is no need to adjust all of our algorithms' transactions. This stage enables us to improve the speed of sanitization and reduce the number of alterations required throughout the hiding process. Additionally, the approach is applicable to both small and big datasets.

An item's existence or absence is recorded on each transaction's chromosome, with a 1 indicating that an item is present, while a 0 indicates that it is not. A chromosome's fitness is governed by a variety of variables and techniques. Each population has many chromosomes, with the best

chromosome being utilized to form the subsequent population. The initial population is made up of a large number of random transactions. The next generation will be shaped by the population's ability to improve survival fitness.

Using various selection processes, GA can guide a population of individuals to an optimal level of "fitness" (i.e., minimizes the cost function). To reduce both lost and ghost rules, the proposed system utilizes two unique fitness functions that alter only transactions with high numbers of sensitive items and small numbers of non-sensitive items. In both scenarios, the transaction with the lowest fitness value will be altered. As stated in [47], the first fitness function f_1 is as follows:

$$\forall t_r \in T \text{ , Compute } \quad f_1 = \frac{X+Y}{2} \tag{7}$$

$$X = \sum_{i=1}^{n}(I_i = 1),\ Y = (S \text{ in } T) \tag{8}$$

$S \in I$ defines the set of sensitive items, T is the set of transactions $T = \{t_r : 1 < r \leq N,\}$, $\mathcal{R}$ defines the set of items $\mathcal{R} = \{I_1, I_2, \ldots\ldots, I_n\}$, n represents the number of items in each transactions. *Y*=(*S* in *T*) indicates the number of items in the transaction that are defined to be sensitive items. In our case, the number of transactions (that is, the maximum value of r) depends on the dataset used in the experiments. For example, the Heart Disease dataset has 303 transactions, while Breast Cancer has 286 transactions. Each dataset has a different number of items. Herein, Instead of using 1s ($I_i = 1$), this fitness function uses 0s ($I_i = 0$) instead of 1s to limit the number of items that may be used. The second fitness function f_2 is based on a weighted sum function and computed as [48]:

$$f_2 = W_1 * C_1 + W_2 * (\frac{1}{C_2}) \tag{9}$$

$$\forall C_1 \in T,\ C_1 = \frac{1}{\sum_1^n Count\ (S)} \quad \text{in } T + \sum_{i=1}^{n} I_i = 1 \tag{10}$$

$$\forall C_2 \in T,\ C_2 = \frac{1}{\sum_1^n Count\ (S)} \quad \text{in } T + \sum_{i=1}^{n} I_i = 0 \tag{11}$$

$$W_1 + W_2 = 1(\ W_1 = W_2 = \frac{1}{2}\ ) \tag{12}$$

Herein, we fixed the weights of W_1 and W_2. Because the system replaces chosen transactions with their offspring that have the most accessible data items, Eq. (10) ensures that lost rules are kept to a minimum, and Eq. (12) ensures that ghost rules are kept to a minimum as well [45-48].

Recently, clustering-based data sanitization approaches were suggested. Cluster mechanisms have a greater impact on the privacy of data in the healthcare sector. The most efficient way of resource allocation with certain restricted conditions among the public, doctors, and medical staff is obtained through the cluster technique. The advantage of the cluster technique is that it efficiently manages the patient's data with access policies, and in this way, privacy preservation is achieved [49]. So, one of the mechanisms through which the proposed framework can operate is the application of the clustering technique as a first step. From this point of view, the work in [50] may represent the basis on which the proposed model is built, in which the fuzzy *K*-medoid machine learning algorithm is used for cluster formation and suggested optimization-based data sanitization is employed for data privacy. The latest references [51-56] provide more information on how machine learning techniques were used for privacy preserving data mining in IoT-based healthcare applications.

IV. EXPERIMENTAL RESULTS

In this section, the suggested technique for hiding interesting negative rules is evaluated and discussed in detail. The experiments use an x64-based processor and 8 GB of DDR3 memory on an Intel ® Core™ i7-5500 M CPU running at 2.50 GHz. MATLAB 7.8.0 was used to create all of the programs. The UCI learning machine laboratory website (http://archive.ics.uci.edu/ml/index.php) provided five healthcare benchmark datasets that were utilized in the studies. The Iris dataset has 150 records and 4 attributes, while the Heart Disease dataset has 303 records and 5 attributes. Breast cancer has 286 records and 9 attributes. The adult dataset has 32,561 records and 10 attributes. Finally, the Diabetes 130-US Hospitals Data Set has 100,000 records with 55 attributes. This data has been prepared to analyze factors related to readmission as well as other outcomes pertaining to patients with diabetes. Emphasis was placed on the attributes related to diabetes (15 attributes).

For implementation, the suggested model was built based on the following GA parameters: The population type is bit strings; the chromosome length fluctuates with the number of items; the population size is 100; the number of generations is 200; the crossover ratio is 0.8; the mutation ratio is 0.1; the fitness function is based on Eq. (7) or Eq. (9); the selection method is a tournament of size two; and the elite count is two. The following measures are used to examine the potential adverse effects of the genetically-based negative association rule concealment approach: There are three ways to cover up mistakes: Hiding Failure (*HF*), Artificial Rule generation (*AR*), and Lost Rules (*LR*) [8] [15].

$$HF = \frac{|R_{sen}(\acute{D})|}{|R_{sen}(D)|} \tag{13}$$

$$AR = \frac{|R\ (\acute{D})| - |R\ (D) \cap R\ (\acute{D})|}{|R\ (\acute{D})|} \tag{14}$$

$$LR = \frac{|R\ _{non-sen}(D) - R\ _{non-sen}(\acute{D})|}{|R\ _{non-sen}(D)|} \tag{15}$$

D and $\acute{D}$ represent the original and sanitized database respectively. $R(D)$ is the set of rules extracted from the original database, while $R(\acute{D})$ is the set of rules extracted from the sanitized database. $R_{sen}(\acute{D})$ represents the set of sensitive rules extracted from the sanitized database, $R_{sen} \subseteq R$, R is the set of all rules. $R_{non-sen} = R - R_{sen}$. Where |.| is the number of items in the collection. Five different sanitization variables were used in the experiments, and the results were averaged. Rule concealment relies on $Min_{sen-conf}$, the number of transactions and the number of items, all three of which are critical inputs. We ran a lot of tests

on each database in order to show how these factors would affect our new system.

A. EXPERIMENTS RELATED TO NEGATIVE RULE MINING EXTRACTION

The suggested method was compared to standard Apriori [20] and genetic-Apriori [29] algorithms for mining negative rules using a variety of input parameters, including support, confidence, and itemset length. The results are summarized in Tables I–III for various healthcare datasets, with the number of interesting negative rules designated by $X \rightarrow \neg Y$ or $\neg X \rightarrow Y$ *or* $\neg X \rightarrow \neg Y$ with 65% support, 55% confidence, and two item length. Notably, since Apriori requires extra runs through the database to generate all the necessary support, it is clearly more time-demanding.

According to the results in the tables, the suggested system mines the fewest negative association rules necessary to attain optimum support and confidence using the fitness function formula. The Tabu's role in condensing the rule search areas inside GA is shown clearly. When the number of database transactions grows, the proposed solution reduces the number of rules by 10% to 50% when the GA algorithm is used alone. We note that the proposed system shows strong performance in the case of large-sized databases with regard to reducing the number of extracted negative association rules, as the importance of the Tabu module appears in reducing the redundant solutions used by the genetic algorithm, and thus only the unique non-repeated rules appear.

TABLE I
TOTAL NUMBER OF GENERATED RULES USING APRIORI ALGORITHM FOR NEGATIVE RULES X →¬Y , ¬ X →Y, AND ¬ X →¬ Y

Datasets	Apriori Algorithm		
	$X \rightarrow \neg Y$	$\neg X \rightarrow Y$	$\neg X \rightarrow \neg Y$
Heart Disease	16	20	14
Breast Cancer	17	17	23
Iris	12	12	16
Adult	2200	3100	2000
Diabetes	5200	6100	4600

TABLE II
TOTAL NUMBER OF GENERATED RULES USING GENETIC-APRIORI ALGORITHM FOR NEGATIVE RULES X →¬Y, ¬ X →Y, AND ¬ X →¬ Y

Datasets	Apriori with Genetic-Apriori Algorithm		
	$X \rightarrow \neg Y$	$\neg X \rightarrow Y$	$\neg X \rightarrow \neg Y$
Heart Disease	7	9	10
Breast Cancer	6	6	8
Iris	5	5	8
Adult	800	1200	1600
Diabetes	2310	2970	3300

TABLE III
TOTAL NUMBER OF GENERATED RULES USING GENETIC-TABU APRIORI ALGORITHM FOR NEGATIVE RULES X →¬Y , ¬ X →Y, AND ¬ X →¬ Y

Datasets	Apriori with Genetic-Tabu Algorithm		
	$X \rightarrow \neg Y$	$\neg X \rightarrow Y$	$\neg X \rightarrow \neg Y$
Heart Disease	6	7	7
Breast Cancer	5	5	7
Iris	4	3	6
Adult	492	396	894
Diabetes	1680	2310	1930

Because of this, a substantial number of rules for negative item sets are no longer retrieved due to the rise in support values and the decrease in confidence values. Similarly, increasing the confidence values in conjunction with the stability of the support value reduces the number of extracted rules, but only by a tiny amount. As a result, it may be stated that support is more effective at reducing the number of mined rules. The time required to extract those rules decreases as the minimum support increases. These results corroborate the conclusions from the surveys' other approaches. Notably, the suggested system requires additional time (a computational cost) to perform Tabu search operations, which contributes to the reduced search space. This results in an improvement in the suggested system's efficiency for extracting negative rules.

The next set of experiments examines the extent to which the proposed model is impacted by many key GA factors, including crossover and mutation rates. To begin with, increasing the ratio of crossover and mutation rates lengthens the time required for GA to extract the rules. In heart disease and breast cancer datasets, we examined various mutation rates ranging from 0% to 20%. The results indicate that although the mutation rate has a little influence, setting it to 0% results in the omission of several interesting rules. A mutation rate of 5-10% is an excellent option since it may provide over 80% of the rules for heart disease and over 90% of the rules for breast cancer. When the mutation rate is reduced by 5%, the average time required to extract a rule is reduced. As a result, we established a mutation rate of 5% for the subsequent experiments.

Similarly, we investigated various crossover rates ranging from 60% to 100% in datasets relating to heart disease and breast cancer. With low crossover rates, such as 60%, we derived almost identical rules to those obtained with high crossover rates. Thus, 60% is the optimal decision for the two datasets used in our studies. For the remaining variables in GA, such as generation number and population size. These factors may have no influence on the Tabu search technique, which is employed to optimize population selection within each iteration.

B. EXPERIMENTS RELATED TO RULE SANITIZATION

For this set of experiments, we're trying to figure out how the number of transactions and the number of hidden negative-sensitive and artificial rules in an adult dataset correlate. In this experiment, $Min_{sup} = 25\%\%$ and $Min_{conf} = 58\%$, while $Min_{sen-conf}$ is set at 60%, 70%, and 80% for 500, 1000, 2000, 3500, and 5000 transactions, respectively. Tables IV and V detail the side effects of the hiding procedure for the fitness functions f_1 (restrictive mode) and f_2 (distortion mode), respectively. As observed in both tables, the loss of non-sensitive rules is quite modest and tends to grow as the number of database transactions rises and fall as the number of sensitive rules$|R_{sen}|$decreases. The suggested approach is determined to have a 0% hiding failure rate, which indicates

that all sensitive rules are protected from exposure. Sensitive rule protection is 100 % accurate.

TABLE IV
PERFORMANCE EVALUATION FOR f_1(%) FOR ADULT DATASET

$Min_{sen-conf}$	60 %			70 %			80 %		
No. of Transactions	*LR*	*HF*	*AR*	*LR*	*HF*	*AR*	*LR*	*HF*	*AR*
1000	0	0	1.28	0	0	1.04	0	0	0.89
2000	0	0	1.42	0	0	1.25	0	0	1.00
3500	0	0	1.70	0	0	1.48	0	0	1.43
5000	0	0	2.13	0	0	2.08	0	0	2.00

TABLE V
PERFORMANCE EVALUATION FOR f_2 (%) FOR ADULT DATASET

$Min_{sen-conf}$	60 %			70 %			80 %		
No. of Transactions	*LR*	*HF*	*AR*	LR	*HF*	*AR*	*LR*	*HF*	AR
1000	0	0.010	1.30	0	0.008	1.05	0	0.005	0.93
2000	0	0.015	1.44	0	0.012	1.29	0	0.010	1.05
3500	0	0.027	1.71	0	0.017	1.59	0	0.014	1.49
5000	0	0.030	2.15	0	0.020	2.13	0	0.018	2.07

As seen in Table V (change of the distortion mode), the number of new rules generated tends to grow in direct proportion to the number of database transactions. We discovered that hiding larger sets of rules results in the introduction of a greater number of new frequent itemsets, which results in the generation of an increasing number of new rules. Unlike the restriction mode modification, which results in zero new rules being mined from the database once the rules are hidden. In other words, adopting f_1 results in improved performance when it comes to ghost rules minimization. However, in both circumstances, the number of transactions that must be adjusted is minimized since the proposed system picks transactions that fulfil maximum modification rules characteristics to alter each time, requiring far fewer transactions to be modified in total.

The next series of experiments compares our technique to that described in [47], which also deals with association rule concealing but does so by hiding itemsets rather than rules, as the proposed system does. The average side effect generated by both systems under the adult database with 5000 transactions covering ten items is shown in Table VII. The results in the table demonstrate that the suggested system missed just a few rules. Additionally, neither approach generated any ghost rules when the specified rules were hidden without causing any side consequences of hiding failure. The illustrations demonstrate that the suggested approach outperforms another technique in terms of reducing side effects and data distortions. As a result, our method had no effect on the quality of the data mining results and took very little time to hide a lot of sensitive association rules from the real database.

TABLE VII
COMPARATIVE RESULTS (%) FOR ADULT DATASET

Algorithm	*LR*	*HF*	*AR*
Proposed Model	2.13	0	0
Comparative model [47]	7.60	0.16	0

The last set of experiments were conducted to validate the efficiency of the suggested model as compared to the traditional *K*-anonymity technique. Herein, the real PIMA Indian diabetes dataset [57] was used as the data input. The datasets consist of several medical predictor variables and one target variable, which is the outcome. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on. Pima Indian Diabetes dataset has 9 attributes in total, and 768 records. The *K*-anonymity technique is run using ARX, which is an efficient open source data anonymization framework that is implemented in Java [58]. The results shown in Fig. 3 confirm the superiority of the proposed model by a 95% improvement with respect to the indicator of hiding failure. Based on the *HF* of the dataset, the results show that the anonymized datasets yield a worse result in the hiding function as the percentage of *HF*, which is around 20 to 22 % for different *k*. Contrary to the proposed model, it achieves 2% *HF*.

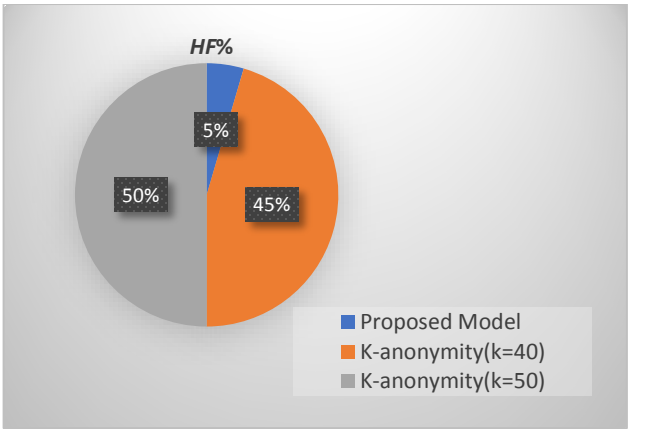

FIGURE 3: Comparative results for PIMA Indian diabetes dataset.

The ultimate goal for all data perturbation algorithms is to optimize the data transformation process by maximizing both data privacy and data utility. However, the two metrics typically represent two conflicting goals in many existing perturbation techniques. Given a data perturbation technique, the higher the level of difficulty in which the original values can be estimated from the perturbed data, the higher the level of data privacy this technique supports. Data utility typically refers to the amount of mining-task/model specific critical information preserved about the dataset after perturbation. The intrinsic correlation between data privacy and data utility raises a number of important issues regarding how to find the right balance between the two measures. The suggested model uses GA to identify the optimum itemsets for alteration, which

is a good transformation that provides a satisfactory level of privacy guarantee. So, it strikes a balance between keeping data useful and accurate and making sure data privacy is protected.

The implementation of differential privacy or local differential privacy on the extracted negative rules may also be used to strengthen the proposed system's privacy. In the context of differential privacy, it is possible to describe the patterns of groups in a dataset publicly while keeping information about individuals. Even if an opponent gets access to an individual's personal replies in the database, they will not be able to learn too much about that person's personal data under a concept of local differential privacy. Table VIII compares the designed model with the state-of-the art (SOTA) models in PPDM.

TABLE VIII
PRIVACY PRESERVING DATA MINING TECHNIQUES ANALYSIS [59][60][62][64]

PPDM Techniques	Anonymization	Perturbation	Randomized	Cryptography	Suggested Model
Advantages	Data owners' sensitive data are to be secured	Preserves various attributes independently	Simple and useful for keeping the sensitive information secretly	Provides better privacy and data utility	Provides better privacy and data utility-Preserves various attributes independently
Limitations	Loss more information, susceptible to attacks	Loss more information, cannot retrieve the original data values	Loss of individual data, it is not suitable for database containing several attributes	Very low scalability	Need more computation to select victim' items
Computation Cost	Low	Low	Low	High	Middle
Privacy Preserving	Average	High	Average	High	High
Mining Accuracy	Average	Average	Average	High	High
Scalability	High	High	Low	Low	High

V. CONCLUSION

Keeping and transferring medical data has become more difficult as concerns about privacy have grown. The spread of healthcare data may be very useful, but it must be done in a way that protects the privacy of patients. It is not an easy task to ensure the privacy of the data that has been disseminated. The healthcare industry has a dilemma in protecting patient data while also making it useful for data mining. There are a number of privacy-protecting data mining methods in use today. The amount of privacy offered by each of these algorithms may be classified in one of three ways: policy-based privacy, statistical privacy, or a combination of the three. As a result, privacy and data use rules for different kinds of healthcare data may be extremely varied.

The purpose of this study is to address the privacy issues associated with healthcare databases as a result of data mining technologies. We used a genetic optimization technique to hide negative sensitive association rules using a heuristic approach based on both distortion and restriction processes. The suggested solution is based on a strategy of concurrently decreasing the confidence in the sensitive rules. The technique makes the fewest possible modifications to the database and misses the fewest possible non-sensitive association rules, which is the ultimate goal of data sanitization. The proposed algorithm is a hybrid of the Apriori and integrated genetic-Tabu algorithms. Rather than mining negative association rules intuitively, the proposed methodology utilizes negative interestingness to describe and explain the success of negative association rules.

By using a genetic-Tabu search method, the system lowers the mining process's search space. The main benefits of the algorithm are that (1) a simple heuristic method is used to choose the transactions and items to be cleaned; (2) a genetic algorithm is used to adjust the victim's choice of items; and (3) data availability is improved by hiding rules instead of items. The fitness function's efficiency has been evaluated in a variety of healthcare databases to determine whether it holds up when a variety of changes are made to the original database. From the simulation results, it is clear that the rules of the suggested technique have much higher support and confidence values while requiring much less processing time to reach the goals.

Privacy-preserving data transformation techniques, log extensions, and process mining algorithms will all be examined in future work, as well as an empirical investigation of how these strategies affect healthcare logs. Furthermore, the suggested method will be tested on a variety of healthcare datasets, including those with varying features, to ensure that it is effective. For more privacy, negative association rules will be applied with differential privacy, local differential privacy, or a combination of both.

REFERENCES

[1] D. Kumari, Y. Vineela, T. Krishna, and B. Kumar, "Analyzing and performing privacy preserving data mining on medical databases", *Indian Journal of Science and Technology*, vol. 9, no. 17, pp. 1-9, 2016.

[2] N. Domadiya, and U. Rao, "Privacy preserving distributed association rule mining approach on vertically partitioned healthcare data", *Procedia Computer Science*, vol. 148, pp. 303-312, 2019.

[3] L. Abuwardih, W. Shatnawi, and A. Aleroud, "Privacy preserving data mining on published data in healthcare: A survey", in *Proc. International Conference on Computer Science and Information Technology*, pp. 1-6, 2016.

[4] A. Pika, T. Wynn, S. Budiono, and A. Hofstede, W. Aalst, H. Reijers. "Towards privacy-preserving process mining in healthcare", In *Proc. International Conference on Business Process Managemen*t, pp. 483-495, 2019.

[5] A. Rashid, and N. Yasin, "Sharing healthcare information based on privacy preservation", *Scientific Research and Essays*, vol. 10, no. 5, pp. 184-195, 2015.

[6] M. Hassan, M. Butt, and M. Zaman, "Privacy preserving data mining for healthcare record: a survey of algorithms", *International Journal of Trend in Scientific Research and Development*, vol. 2, no. 1, pp. 1176-1184, 2017.

[7] A. Pika, T. Wynn, S. Budiono, "Privacy-preserving process mining in healthcare ", *International journal of environmental research and public health*, vol. 17, no. 5, pp.1-28, 2020.

[8] K. Oishi, Y. Sei, Y. Tahara, and A. Ohsuga, "Semantic diversity: Privacy considering distance between values of sensitive attribute", Computers & Security, vol. 94, 101823. pp.1-17, 2020.

[9] S. Darwish, M. Madbouly, and M. El-Hakeem, "A database sanitizing algorithm for hiding sensitive multi-level association rule mining", *International Journal of Computer and Communication Engineering*, vol. 3, no. 4, pp. 285, 2014.

[10] M. Dehkordi, K. Badie, and A. Zadeh, "A novel method for privacy preserving in association rule based on genetic algorithms," *Journal of software*, vol. 4, no. 6, pp. 555-562, 2009.

[11] R. Crawford, M. Bishop, B. Bhumiratana, L. Clark, and K. Levitt "Sanitization models and their limitations", In *Proc. of The Workshop on New Security Paradigms*, pp. 41-56, 2006.

[12] H. Waguih, S. Darwish, M. Osman, "Mining interesting positive and negative association rule based on genetic Tabu heuristic search", *Journal of Theoretical and Applied Information Technology*, vol. 96, no. 23, pp. 7834-7845, 2018.

[13] J. Lin, J. Wu, P. Fournier-Viger, Y. Djenouri, C. Chen, Y. Zhang, "A sanitization approach to secure shared data in an IoT environment", *IEEE Access*, vol. 7, pp. 25359-25368, 2019.

[14] D. Toshniwal, "Privacy preserving data mining techniques for hiding sensitive data: a step towards open data", *Data Science Landscape*, pp. 205–212, 2018.

[15] V. Verykios, E. Bertino, I. Fovino, L. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining", ACM Sigmod Record, vol. 33, no. 1, pp. 50–57, 2004.

[16] E. Dasseni, V. Verykios, A. Elmagarmid, and E. Bertino, "Hiding association rules by using confidence and support,'' In *Proc. International Workshop on Information Hiding*, pp. 369–383, 2001.

[17] R. Oliveira, and O. Zaiane, "Privacy preserving frequent itemset mining", In *Proc. IEEE International Conference on Privacy, Secure Data Mining*, pp. 43–54, 2002.

[18] M. Islam, and L. Brankovic, "Privacy preserving data mining: a noise addition framework using a novel clustering technique", *Knowledge Based System*, vol. 24, no. 8, pp. 1214–1223, 2011.

[19] X. Liu, S. Wen, and W. Zuo," Effective sanitization approaches to protect sensitive knowledge in high-utility itemset mining" *Applied Intelligence*, vol. 50, no. 1, pp. 169-191, 2020.

[20] X. Sun, and S. Philip, "Hiding sensitive frequent itemsets by a border-based approach", *Journal of Computing Science and Engineering* vol. 1, no. 1, pp. 74-94, 2007.

[21] M. George, and S. Verykios, "A MaxMin approach for hiding frequent itemsets", *Data & Knowledge Engineering*, vol. 65, no. 1, pp.75-89, 2008.

[22] A. Amiri, "Dare to share: protecting sensitive knowledge with data sanitization", *Decision Support Systems*, vol. 43, no. 1, pp.181-191, 2007.

[23] Wang, S. Liang, B. Parikh, and A. Jafari, "Hiding informative association rule sets", *Expert Systems with Applications*, vol. 33, no. 2, pp. 316-323, 2007.

[24] Y. Wu, C. Chiang, and A. Chen, "Hiding sensitive association rules with limited side effects", *IEEE Transactions on Knowledge and Data engineering*, vol. 19, no. 1, pp. 29-42, 2006.

[25] A. Gkoulalas, V. Verykios, "Exact knowledge hiding through database extension". IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 5, pp. 699-713, 2008.

[26] C. Wu, and Y. Huang, "A cost-efficient and versatile sanitizing algorithm by using a greedy approach", *Soft Computing*, vol. 15, no. 5, pp. 939-952, 2011.

[27] P. Cheng, I. Lee, C. Lin, and J. Roddick, "Hide association rules with fewer side effects", *IEICE Transactions on Information and Systems*, vol. 98, no. 10, pp. 1788-1798, 2015.

[28] T. Hong, C. Lin, K. Yang, and S. Wang, "Using TF-IDF to hide sensitive itemsets", *Applied Intelligence*, vol. 38, no. 4, pp. 502-510, 2013.

[29] C. Wei, T. Hong, J. Wong, G. Lan, and W. Lin, "A GA-based approach to hide sensitive high utility itemsets", *the Scientific World Journal*, vol. 2014, Article ID 804629, pp.1-13, 2014.

[30] U. Yun, and J. Kim, "A fast perturbation algorithm using tree structure for privacy preserving utility mining", *Expert Systems with Applications*, vol. 42, no. 3, pp. 1149-1165, 2015.

[31] T. Wu, J. Zhan, and J. Lin, "Ant colony system sanitization approach to hiding sensitive itemsets", *IEEE Access*, vol. 5, pp. 10024-10039, 2017.

[32] T. Wu, J. Lin, Y. Zhang, and C. Chen, "A grid-based swarm intelligence algorithm for privacy-preserving data mining", *Applied Sciences*, vol. 9, no. 4, 774,pp.1-20, 2019.

[33] A. Divanis, and V. Verykios, "An overview of privacy preserving data mining", *The ACM Magazine for Students*, vol. 15, no. 4, pp. 23-26, 2009.

[34] L. Lekshmy, and A. Rahiman, "A sanitization approach for privacy preserving data mining on social distributed environment", *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 7, pp. 2761-2777, 2020.

[35] J. Wu, G. Srivastava, A. Jolfaei, P Fournier-Viger, J. Lin," Hiding sensitive information in eHealth datasets", *Future Generation Computer Systems*, vol. 117, pp. 169-180, 2021.

[36] J. Wu, G. Srivastava, U. Yun, S. Tayeb, and J. Lin. "An evolutionary computation-based privacy-preserving data mining model under a multithreshold constraint", *Transactions on Emerging Telecommunications Technologies*, vol.32, issue 3, e4209, pp: 1-19, 2021.

[37] A. Hasan, Q. Jiang, J. Luo, C. Li, and L Chen," An effective value swapping method for privacy preserving data publishing", *Security and Communication Networks*, Vol. 9, no. 16, pp. 3219-3228, 2016.

[38] A. Zigomitros, F. Casino, A. Solanas, and C. Patsakis," A survey on privacy properties for data publishing of relational data", *IEEE Access*, vol.8, pp. 51071-51099, 2020.

[39] K. Ranjith, and A. Geetha Mary," Privacy-preserving data mining in spatiotemporal databases based on mining negative association rules", In *Emerging Research in Data Engineering Systems and Computer Communications*, pp. 329-339, 2020.

[40] A. Telikani, and A. Shahbahrami, "Data sanitization in association rule mining: an analytical review", *Expert Systems with Applications*, vol. 96, pp. 406-26, 2018.

[41] J. Lin, T. Wu, P. Fournier-Viger, G. Lin, T. Hong, and J. Pan, " A sanitization approach of privacy preserving utility mining." In *Proc. International Conference on Genetic and Evolutionary Computing*, pp. 47-57. Springer, Cham, 2015.

[42] S. Bagui, and P. Dhar," Positive and negative association rule mining in Hadoop's MapReduce environment", *Journal of Big Data*, 6(1):1-6, 2019.

[43] A. Kadir, A. Bakar, and A. Hamdan," Frequent absence and presence itemset for negative association rule mining", In *Proc. 11th International Conference on Intelligent Systems Design and Applications*, pp. 965-970, 2011.

[44] C. Cornelis, P. Yan, X. Zhang, and G. Chen," Mining positive and negative association rules from large databases", In *Proc. IEEE Conference on Cybernetics and Intelligent Systems*, pp. 1-6, 2011.
[45] N. Rai, S. Jain, and A. Jain," Mining interesting positive and negative association rule based on improved genetic algorithm (MIPNAR_GA)", International Journal of Advanced Computer Science and Applications, 5(1):1-10, 2014.
[46] N. Rai, S. Jain, and A. Jain, "mining positive and negative association rule from frequent and infrequent pattern based on imlms_ga", International Journal of Computer Applications, 77(14):48-52, 2013.
[47] S. Narmadha and S. Vijayarani, "Protecting sensitive association rules in privacy preserving data mining using genetic algorithms," International Journal of Computer Applications, vol. 33, no. 7, pp. 37-34, 2011.
[48] P. RajyaLakshmi, C. M. Rao, M. Dabbiru, and K. V. Kumar, "Sensitive itemset hiding in multi-level association rule mining," International Journal of Computer Science & Information Technology, vol. 2, no. 5, pp. 2124-2126, 2011.
[49] F. Ullah, I. Ullah, A. Khan, M. Uddin, H. Alyami, and W. Alosaimi, "Enabling clustering for privacy-aware data dissemination based on medical healthcare-IoTS (MH-IoTS) for wireless body area network", *Journal of Healthcare Engineering*, vol. 2020, Article ID 8824907, pp.1-10, 2020.
[50] M. Madbouly, S. Darwish, N. Bagi, and M. Osman," Clustering big data based on distributed fuzzy k-medoids: an application to geospatial informatics", *IEEE Access*, vol. 10, pp. 20926- 20936, 2022.
[51] U. Ahmed, G. Srivastava, and J. Lin," A machine learning model for data sanitization", *Computer Networks*, vol. 189:107914, pp.1-6, 2021.
[52] J. Wu, G. Srivastava, M. Pirouz and J. Lin," A GA-based data sanitization for hiding sensitive information with multi-thresholds constraint," In Proc. of the International Conference on Pervasive Artificial Intelligence, pp. 29-34, 2020.
[53] A. Khedr, W. Osamy, A. Salim, and A. Salem," Privacy preserving data mining approach for IoT based WSN in smart city. *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 8, pp. 555-563, 2019.
[54] R. Lu, K. Heung, A.Lashkari, and A. Ghorbani," A lightweight privacy-preserving data aggregation scheme for fog computing-enhanced IoT", *IEEE access*, vol. 5, pp.3302-3312, 2017.
[55] H. Li, F. Guo, W. Zhang, J. Wang, and J Xing," (a, k)-Anonymous scheme for privacy-preserving data collection in IoT-based healthcare services systems", *Journal of Medical Systems*. vol. 42, no. 3, pp.1-9, 2018.
[56] J. Du, C. Jiang, E. Gelenbe, L. Xu, J. Li, and Y. Ren," Distributed data privacy preservation in IoT applications", *IEEE Wireless Communications*. Vol.25, no.6, pp. 68-76, 2018.
[57] https://archive.ics.uci.edu/ml/datasets.php
[58] https://arx.deidentifier.org/development/framework/
[59] S.Shimona, " Survey on privacy preservation technique", In *Proceedings of the International Conference on Inventive Computation Technologies*, pp. 64-68, 2020.
[60] Y. Yang, and Y. Zhou," A survey on privacy-preserving data mining methods. In *IOP Conference Series: Materials Science and Engineering*, vol. 782, no. 2, pp.1-10, 2020.
[61] P. Shynu, H. Shayan, and C. Chowdhary," A fuzzy based data perturbation technique for privacy preserved data mining", In *Proceeding of the International Conference on Emerging Trends in Information Technology and Engineering*, pp. 1-4, 2020.
[62] N. Nasiri, and M. Keyvanpour, " Classification and evaluation of privacy preserving data mining methods", In *Proceedings of the International Conference on Information and Knowledge Technology*, pp. 17-22, 2020.
[63] P. Lekshmy, and M. Rahiman, " A sanitization approach for privacy preserving data mining on social distributed environment.", *Journal of Ambient Intelligence and Humanized Computing*, vol.11, no. 7, pp. 2761-2177, 2020.
[64] S. Rathod, and D. Patel, "Survey on privacy preserving data mining techniques", *International Journal of Engineering Research & Technology*, vol. 9, no. 6, pp. 832-839, 2020.

SAAD M. DARWISH received the B.Sc. degree in Statistics and Computer Science from the Faculty of Science, Alexandria University, Egypt in 1995. He held the M.Sc. degree in information technology from the Institute of Graduate Studies and Research (IGSR), Department of Information Technology, University of Alexandria in 2002. He received his Ph.D. degree from the Alexandria University for a thesis in image mining and image description technologies. He is the author or coauthor of 50+ papers publications in prestigious journals and top international conferences and also received several citations. He has served as a Reviewer for several international journals and conferences. He has supervised around 90 M.sc and Ph.D. students. His research and professional interests include image processing, optimization techniques, security technologies, database management, machine learning, biometrics, digital forensics, and bioinformatics. Since June 2017, he has been a professor in the department of information technology, IGSR.

REHAM M. ESSA received the B.Sc. degree in quality education, Department Technology Education from the Faculty of Specific Education, Tanta University, Kafr El-Sheikh Branch, Egypt in 2003. She held the M.Sc. degree in Specific Education from Tanta University. She received his Ph.D. degree from the Cairo University. She has served as a reviewer for several international journals and conferences. Her research and professional interests include data mining, optimization techniques, database management, machine learning. Since 2020, she has been associate professor in the department of information system and computer sciences.

MOHAMED A. OSMAN received the B.Sc. degree in Management Information system from the Higher Institute of Management and Information Technology, Kafr El-Sheikh, Egypt in 2009. He held the M.Sc. degree in computer and information systems from Sadat Academy for Management Sciences, Department of Information Systems, Faculty of Management Sciences in 2019. His research and professional interests include optimization techniques, security technologies, and database management.

Ahmed A. Ismail received the B.Sc. degree in Management Information System (MIS) from the College of Business Administration, Arab Academy for Science & Technology and Maritime Transport in 1999. He held the M.Sc. degree in information technology from the Institute of Graduate Studies and Research (IGSR), Department of Information Technology, University of Alexandria in 2007. He received his Ph.D. degree from Helwan University for a thesis in tracing and detecting explosives. He has more than 15 year experience in the field of Information Technology. His research and professional interests include e-marketing, business intelligence, business information systems, management information systems, database management systems, web programming, data communication and computer networks.