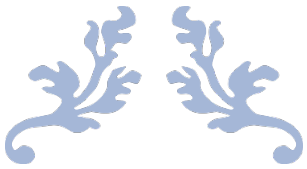


GENETIC ALGORITHM BASED PRIVACY PRESERVING AND EFFICIENT RULE GENERATION IN DISTRIBUTED ENVIRONMENT

Safeer Ahmed

Bachelor Thesis
Department of Informatics

Acknowledgment

This bachelor thesis is the final project for a Bachelor degree in Network Technologies at New Bulgarian University.

I would like to thank my supervisor Rossitza Ivanova Goleva for giving me an opportunity to dive into such a complex investigation and for her support and brain-storming. I could not have asked better person to support me through this long and tedious journey.

Also I'm extremely happy to have my close friends and classmates which I can share ideas and dig into psychophysical conversations with. Finally, I am proud to have such an amazing family which have given me the support, love and courage to complete this task.

New Bulgarian university, October, 2021

Safeer Ahmed

Abstract

The world has become more digitalized with the rapid growth in science and technology, with that the growth of data becomes higher and it brought an essence of data storing and data privacy. It has become a huge task to extract useful information from raw data without breach of privacy and damage the actual data. Models and algorithms were produced for ensuring the benefit and privacy of data are categorized in the specialized field of data mining term as Privacy-Preserving Data Mining (PPDM). The necessity of privacy technique k-anonymity for posting microdata demands that every single uniform class (for instance, a set of records that are identical to each other concerning certain "identifying" traits) contains minimum k records. As of late, a few creators have acknowledged that k-anonymity cannot avoid attribute divulgence. The perception of ℓ-diversity has been suggested to convey this; ℓ-diversity needs that every single uniform class has minimum ℓ well-represented values for all subtle attributes. Lately, many authors have understood that ℓ-diversity cannot avoid probabilistic deduction attacks. To talk about the restriction of ℓ-diversity, t-closeness privacy technique has been proposed; t-closeness demands that the distribution of a sensitive attribute in an equivalence class is close to the arrangement of the attribute in the whole table (for example, there should be no more than a threshold t distance between two distribution). T-closeness privacy model uses a fixed threshold t which cannot issue the arrangement of a sensitive attribute in an equivalence class adjacent to the arrangement of the attribute in the whole table. This reduces the privacy of the data. To address this issue, a framework is developed to optimize the threshold t of t-closeness using the Genetic Algorithm (GA) and provide the best optimal solution according to the provided data sets and address the objective of maximizing the fitness of the data and privacy. GA is used for the optimization process as they are easily adaptable and applicable. With this method, the threshold value of t-closeness is easily computable and optimized. By using the GA optimized t-closeness, the usefulness of data and its privacy can be maximized. The classification process is based on multiple learning classifiers to provide more accurate and efficient rules for higher classification accuracy. To verify the improved effectiveness and utility of produced results, experimentation is performed using the percentage split on different datasets.

Table of Contents

List of Tables

Table of Figures

Chapter 1

Introduction

1 Introduction

Past few years, stupendous progress has been detected in the quantity and variability of information related to a specific subject which is collected from several parties that participate in a common task, for example; medical researchers collect data from different hospitals for research purposes. On the other side, severe threats are positioned to the secrecy of sensitive data. In the context of protection laws and regulation authorities, preserving the privacy of sensitive data is one of the necessary aspects. These laws are made for protecting the sensitive data is published and set down an agreement between the participants before the sensitive data is published and analyzed, for instance, hospitals and clinics keep patient history like demographic information of patients, their payments records, vaccination, and laboratory test results. The Patient trusts their doctors and the concerned medical authorities not to expose or share their personal information for any purpose with a third party. However, for disease treatment and prevention research, medical research needs subject-specific sensitive information from hospitals. A challenge arises, when trying to preserve company's commercial secrets, or preserve the privacy of a single and client by exchanging shared data or information uses for analytical objectives and protecting that information by keeping it private. Several developing technologies, for example, cellular phone and Radio-Frequency Identification (RIFD) tags provide information based on location, digital cameras taking pictures, the web provides GPS or patterns for navigation, online transactions for e-commerce, display ads by e-mail scanning, networking, and so on. Which has made very important privacy and secrecy measures to the point where inescapable situations are preceded. However, every person must have fundamental authority about what type of data about them is being collected? Duration for storing that data? Who will have access to that data? For what purpose it will be used?

Within the field of data mining to preserve privacy, several techniques and algorithms have been developed in the last few years. The representation granularity is decreased by incorporating privacy in the existing techniques. Such decreases are the reason for the misfortune of adequacy within the application comes created from the data mining algorithms. The privacy obstacles have been argued and addressed in different concerned communities, for example, database experts, statistical disclosure control specialists, and the cryptography community. Let's take an example, a girl named Helena shared her private information like a home address and phone number in the directory. She gave her marital status and her age when she was ill and admitted to the hospital. In one of her social media accounts, Twitter, she tweeted that 'Hey friends, I am

in Varna with family and enjoying my holidays'. Adversary got to know about all these situations and by relating all of them together, he/she gets to know that who is Helena and where she lives and other information related to it.

Evolutionary Algorithms (EAs) are population-based speculative search algorithms that are motivated by the method of Darwinian evolution [1]. In a developed technique genetic algorithm is used to optimize the threshold value of t-closeness. Genetic Algorithm (GA) is one of the well-known and recognized algorithms of EA [2]. GA is a population-based metaheuristic search algorithm, which is used to resolve optimization problems. GA is used for the developed technique because in the data privacy-preserving field its capabilities have been recognized as a machine learning technique. Using this method, we can enhance the ability to effectively prevent privacy leaks, and ensuring the availability of data quality.

1.1 Problem Statement

- The existing privacy approach t-closeness demands that the arrangement of an intuitive attribute in a uniform class is close to the distribution of the attribute in the whole table (for example, there should not be more than a threshold t distance between two distributions). T-closeness privacy model uses a fixed threshold t which cannot provide an arrangement of a sensitive attribute in a uniform class close to the distribution of the traits in the whole table. It degrades the data utility and data privacy.
- No standard strategy to authorize t-closeness.

1.2 Research Question

- What measurements should be taken to optimize the threshold (t) value of t-closeness to increase the privacy of data without degrading the data utility?
- What will be the optimal solution to generate more accurate classification rules?

1.3 Research Objective

The primary aim of this research work is to address the concerns related to privacy of data. Optimizing the value of t-closeness to increase the privacy of data and generate more accurate and efficient classification rules to produce more accurate results.

1.4 Research Contribution

In the privacy-preserving data-mining (PPDM) area of investigation, there're two possible methods, in which privacy and accuracy both are considered together in group anonymization techniques. The first method is the Mine-and-Anonymize approach, in this approach, initially the classification is applied to the original data set via traditional data mining methods, and after that anonymization techniques are applied to the results of classification to preserve the privacy of sensitive information. The second method is the Anonymize-and-Mine approach, in this approach, initially anonymization techniques are applied to the original data before passing them to the classification method, and later classification is applied on the anonymized data for getting the classification results. There's a dilemma between the precision of results and privacy metrics in both approaches. According to the required amount of applicability, generalization, and distortion of data also needs to be controlled. In this thesis, a framework is developed to address the goal to maximize the usefulness of data and privacy in a controlled environment. This framework consists of different processes from generating classification rules from multiple classifiers j48, ripper, and part-based learning algorithms [3] [4] [5]. The reason behind using rules generated from multiple classifiers is that it generates better classification results as compared to results generated by a single classifier. Here a question arises regarding multiple classifiers that which classifiers should be selected for a given situation to form an optimal group. Also, the computational cost of an ensemble is often very high because they require the execution of multiple classifiers for a single classification task. These problems are addressed in the developed technique by producing a hybrid approach that is easy to implement, and its computational cost is low. In the developed approach for classification, learning classifiers are selected according to the dataset type and working, then combining rules generated from multiple classifiers in a single place. The Effectiveness and efficiency of the developed work are shown by the experimental results. Calculating Intra and intercross similarities. Optimizing the threshold value t of t-closeness using GA to increase the diversity of the released tables by equally distributing sensitive traits in an equivalence class and the whole table. Classification rules are grouped via clustering techniques based on the distance calculated between attributes/rules. Once similar classification rules are grouped in the buckets, instances are mapped to the rules accordingly. In the final process, sensitive rules and instances are filtered out from the buckets to perform the generalization method according to the selected group

anonymization technique. The developed technique provides enough flexibility to get compatible with cryptographic methods.

1.5 Layout of Thesis

- **Chapter 1** gives a basic introduction to the research topic and thesis layout.
- **Chapter 2** describes all the background knowledge in-depth to understand the developed algorithm. It includes basic information about Privacy-preserving Data-mining (PPDM), a Genetic Algorithm, and an ensemble of classifiers.
- **Chapter 3** describes the developed technique in detail. It focuses on primary and secondary research contributions. The principles, methods, and algorithms are developed and implemented to increase the level of privacy of the provided data sets and classification rules generated from those data sets and generate accurate and efficient rules for best classification results.
- **Chapter 4** provides experimental results as the output of the newly developed solution.
- **Chapter 5** concludes the research and also suggests some future work.

Chapter 2

Background Study & Literature Review

2. Background Study and Literature Review

Data mining (DM) is a vast field of research. In this chapter, the background study and literature review of three main research domains of data mining related to the developed framework are discussed. These domains are privacy-preserving data mining (PPDM) techniques, genetic algorithm (GA) in the data mining field, and ensemble learning of multiple classifiers in data mining.

2.1 Privacy Preserving Data Mining (PPDM)

In the past few years, the privacy-preserving data mining problem has become more significant because of the growing capacity for storing personal data and information. Especially, the recent developments in the field of data mining have directed to risen the business related to privacy. Many data mining researchers are focusing on either transforming data or reducing data granularity. The most important legal framework is the European General Data Protection Regulation (GDPR), which become active in 2018, which consists of ninety-nine (99) articles and one hundred and seventy-three (173) interpretative explanations. Some of the principles are lawfulness, transparency, purpose limitation, data minimization, the accuracy of processed data, limitation of duration, integrity, etc.[62] There is a horse-trading between downgrading the data and privacy of the data. Many techniques have been proposed in the last couple of years to execute privacy-preserving data mining. Moreover, this issue has discoursed in various data mining communities (database, cryptography, and statistical disclosure control).

Following are different research areas in this field [6] [7]:

- ### Application Results preservation

  In several situations, the outcomes of data mining techniques; for example, association rule mining (ARM) or classification run the show mining can compromise the protection of the data. This area of research, covers hiding the results of the data mining algorithms, which are helpful for the attackers to foresee the original data from the revealed semantics of data. To assure privacy and decreasing the damage of useful information to its minimal, modifications are required in the existing mining Techniques.

- Query Auditing Privacy Preserving

Query auditing privacy-preserving is the same as the 'Application Results Preservation' approach. This area of research is entirely dependent on restricting and modifying the results generated from the queries. This research area protects the change of the in-flows and out-flows of data.

- Cryptographical Technique for Distributed Privacy

In several cases, the data is distributed with many sites and it requires a mutual function for communication between multiple sites. In such situations, various cryptographic rules may be used in such situation to provide secure communication among the different sites without disclosing sensitive information.

2.1.1 Randomization Technique

The Randomized response is a statistical method that was presented by Warner for solving a survey problem [8]. In the Randomization technique [9], information is twisted by adding some data noise to the original data to make the values of the attributes hidden from the datasets. The added data noise should be huge enough that the original data can't be regained. Then, algorithms are constructed to originate the aggregate distribution from the altered data records. Afterward, to cooperate with these aggregate distributions' data mining algorithms can be constructed. This practice is very effortless and it doesn't require any information on the distribution of the other values in the data. Thus, the randomization approach may be carried out at the time of the facts series. The Randomization method doesn't require a depended on the server to incorporate all of the original statistics to carry out the system of anonymization. The susceptible point of a randomization reaction-based privacy-preserving facts mining method is that it copes with all of the information equally irrespective of their local density. This ends in a problem wherein the outlier facts emerge as more vulnerable to opposed assaults compared to facts in extra dense areas inside the statistics [10]. Concerning this one answer is to be unnecessarily greater competitive in inserting noise into the information. However then, it decreases the application of the records for mining purposes because the recreated distribution won't generate effects according to the purpose of data mining.

The randomization method is explained below:

Suppose that X = { x1,….,xn } is a set of data records, such that xi ∈ X. Add noise to X which is obtained from the probability distribution $f_y(y)$. The noise components are represented as: Y = { y1,….,yn } and they are selected individualistically. As a result, the new deformed set of record is represented as: Z = { x1 + y1,….,xn + yn } represent this set as: Z = { z1,….,zn }. It is presumed generally that the noise inserted into a dataset should be huge enough that the original data records can't be effortlessly restored from the deformed data. Therefore, the original data values can't be retrieved. But the original data records distribution can be retrieved if a random variable X represents the original data record distribution, a random variable Y describes the noise distribution, and a random variable Z denotes the resulting records.

The formal notation of the above-described technique is:

$$Z = X + Y \qquad (1)$$

$$X = Z - Y \qquad (2)$$

Noise can be added to the data through different randomization methods [11]. With the randomization method, the following perturbation approaches are possible:

- Additive Perturbation

In this approach of randomization, a randomized noise is attached to the data values to distort the original data. From randomized data values, the entire data distribution can be regained. Data mining methods are required to be upgraded to utilize randomized data for mining purposes.

- Multiplicative Perturbation

This is a common method of randomization. In the multiplicative perturbation method, random projection or random rotation methods are used to perturbing the data values. This method preserves the bury-record separations quite closely and according to the distorted records which could be used in combination with a combination of data-mining approaches. Two methods are to be recognized in the multiplicative perturbation technique. Predominant

methodology adds noise to enlarge the original data with mean μ=1 and low variance. Another method is first to calculate the log-transformed data, after that processing the covariance lattice the data which is altered generating random numbers. These random numbers have mean μ=0 and earlier considered variance's variance for 'y' times. In the distorted form of data, the noise will be added. Finally, antilog will be calculated for noisy data [12].

- Data Swapping

The data swapping method was introduced by Resis and Dalenius[13] for statistical disclosure protection in private databanks. In the data swapping technique [14], the values of different records are swapped with each other to do privacy preservation. The advantage data swapping method is that the lesser order minimal sums of the data are protected and not concerned. So, with no violation of data privacy, some sorts of combined calculations can be done. This method doesn't stick to the overall principle in randomization that permits the record value to be perturbed individualistically of other records. For that reason, this method can be used with other privacy-preserving frameworks for example k-anonymity, provided that the swapping technique is intended proposed to maintain the definitions of privacy for k-anonymity.

2.1.2 Group Base Anonymization

The Randomization technique is very simple because it can be executed effortlessly at the time of data gathering, this is because the noise appended is liberated of the behavior of original information/data. The disadvantage of the randomization method is the outlier records might be hard to mask. In such cases where at the time of data collection, the privacy-preserving is not required to be performed, for such situations, a technique is required in which the amount of miscalculation is reliant on the manners of the locality of the given tuple. Randomization technique has another drawback, that it does not count the chance that the records which are available publicly, the identity of the owner of that record can be disclosed using that record. Records that are available publicly, their privacy is compromised in case of high dimensional data. To solve such problems, a group-based anonymization method was proposed which constructs groups with anonymous records [7].

In the group-based anonymization method, a person's identity is distorted or completely hide in such a way that the person gets concealed in the crowd. The records of data that are published, don't have any key identifiers (like a person's name, NIC number, contact information) that identify a record distinctively using personal information. Another type of identifier is pseudo identifiers, these attributes can be used to identify the record using information that is available somewhere else, i.e. adversary can identify the particular record in the data set of patients using external information which he got from bank data. Pseudo records are like: DOB, sex, state, etc. by grouping such attributes, the possibilities discloser of identities of individuals, can be overcome.

"Let RT (Released Table) (A1,….,An) may be a table and QI_{RT} (Quasi Identifiers (A_i,.…,A_j)) be the quasi-identifier related with it. RT is said to fulfill k-anonymity if as it were that each arrangement of values in RT [QI_{RT}] shows up with at slightest k events in RT [QI_{RT}]" [15].

Attributes Set = (A_1,…..,A_n)

2.1.2.1 K-Anonymity Method

The k-anonymity technique protecting the released data against the re-identification of records of the particular individuals. It is said to be the table is k-anonymous only if, every single row in the data table is indistinguishable from the rest of the k-1 rows by just viewing its properties. I.e.: If someone tries to identify an individual from a table, one must-have information like DOB, sex, about that particular individual. In the table, k individuals fulfill the requirement [15]. Data is made publicly available in many applications by just eliminating or hiding key identifiers like names, NIC numbers, etc. Though, other types of attributes can be used to correctly recognize the particular records. Such attributes are called pseudo-identifiers and these are like: age, area code, sex, etc. pseudo-identifiers are easily available in community records for example census data. Such records can be used to disclose the identity of that particular individual. The possibilities of identification of a large number of individuals from a released data can be limited by grouping pseudo-identifiers.

In k-anonymity, the graininess of the presentation of pseudo-identifiers can be reduced by using generalization and suppression techniques.

- Generalization

In this technique, the values of the attributes are generalized to a run in arrange to diminish the graininess of representation. I.e., age could be generalized to a range for example set a range 20-29, to decrease the possibility of identity disclosure [14].

Table 1: Original Released Table RT (Micro Data)

Age	Gender	Postal-Code	Salary
32	male	44000	63k
24	Male	44000	50k
29	Female	31000	45k

Table 2: Anonymous Version of Released Table (Age is Generalized)

Age	Gender	Postal-Code	Salary
3*	male	44000	63k
2*	Male	44000	50k
2*	Female	31000	45k

- Suppression

In this technique, the attribute value is removed. The attributes which do not effect anonymization results are removed completely from the data. By suppressing such attributes, it does not effect on the results of the data [14].

Table 3: Anonymous table (gender attribute suppressed)

Age	Gender	Postal-Code	Salary
32	*	44000	63k
24	*	44000	50k
29	*	31000	45k

Suppression and generalization methods decrease the risk of identification of individual records which are publicly available. However, such methods reduce the accuracy of the results generated from distorted data.

2.1.2.2 Approximation Algorithm

In this algorithm, from the actual PT (Private Table), a complete weighted graph is built. In that graph, every one vertex relates to a data record in a PT (Private Table). The edges of the graph are weighted with the number of diverse property values between the two tuples indicated by the most distant vertices. Afterward, the approximation algorithm then constructs a forest of trees that had at as a minimum k vertex signifying the clusters for k-anonymity. Few fields in the vertices are repressed to get that every single record in the alike tree having the similar values of quasi-identifier. The price of a vertex is calculated from the number of repressed cells, the aggregate of the weights of vertices of the tree is the cost of the tree. The cost of the concluding solution is equivalent to the aggregate of the costs of the trees. In building the forest of trees, the maximum possible sum of vertices in a tree which is limited by the approximation algorithm [7] is 3k−3. Without raising the final solution cost, all partitions which have greater than 3k − 3 elements are decayed. Trees not having more than 3k – 3 vertices, guarantees a solution with O(k)-approximate.

2.1.2.3 Condensation Approach

Managing string data is very challenging because of variant lengths in different records. Generally, an attribute that creates difficulty to anonymize records of variant length, the k-anonymity method is more definite about those attributes. A Condensation approach is suggested to anonymize such sorts of string data. This approach creates clusters from such unique strings. After that, it creates distinctive clusters as faux data. There must be at least k-records in each cluster to satisfy the definition of k-anonymity so it will ensure that the data is anonymized [7].

The Condensation approach [16] follows a procedure in which it condenses the data into various groups of predefined size. A particular level of statistical information regarding distinctive records are preserved for all groups. This statistical information is sufficient enough to maintain info regarding the correlations and mean over various dimensions. It is not possible to differentiate various records from each other inside a group. Every group has a particular k of

smallest size which points to the level of indistinguishability of some privacy-preserving method. Larger the indistinguishability point, the larger the privacy measure. Simultaneously, since of the condensation of a better number of records into a single statistical group unit, a huge volume of data is lost.

It must make sure that anonymity of original private table PT shouldn't be despoiled in the k-anonymity technique. To guarantee k-anonymity for data mining, the two possible approaches are:

- Anonymize-and-Mine

  First, apply the k-anonymity technique to anonymize the original private table PT, and an anonymized table PT_k is released which is a k-anonymous version of the original table PT. After anonymization is done, data mining techniques are applied on anonymized released table PT_k by data holders or external parties.

  Figure 1 provides a graphical representation of the anonymize-and-graphical approach.

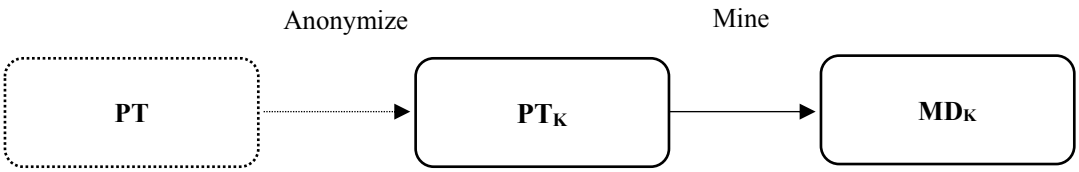


Figure 1 Anonymize-and-Mine Approach **[7]**

Advantages of the Anonymize-and-Mine approach are:

- ✓ Data mining is performed on the anonymized version of private table PT_K so this approach guarantees that data mining is safe.
- ✓ This approach allows external parties other than the data holders to perform data mining so it may help the recipients, they can analyze data and classify the anonymized when the owner of data may not know a priori.
- ✓ The recipients can specifically characterize parameters, for instance, precision and interpretability using application-specific data mining algorithms which they already have and they can also decide the procedures of data mining after examining the data.

On the other side, there are some disadvantages of the Anonymize-and-Mine approach:

- ✓ Since data mining is performed on anonymized data, so utility and importance of the results of the mining process could be compromised. Anonymization uses a k-anonymity algorithm which may generate results that aren't suitable for data mining because k-anonymity methods may conceal the information which is very much beneficial for data mining.
- ✓ This approach may not be beneficial when the information may be only acquired once for example, when the information (data) source is a stream.
- ✓ Anonymize-and-Mine approach is less efficient and expensive in the can of larger and sparse data sets.

- ## Mine-and-Anonymize

In this approach, first, use data mining techniques on the original private table PT. After mining is done, the anonymization technique k-anonymity is applied to mine data to anonymize it. In this approach, data mining can only be executed by data holders. Only the mined data MD_K is delivered to external groups. This approach uses two methods for anonymization; one-step mine-and-anonymization and two-step mine-and-anonymization.
Only the data holders can execute the data mining process. External parties get mined version of the original table MD_K. Usually, impossible to allow any derivation to breach k-anonymity for the original PT. The original private table PT doesn't need to be k-anonymous, and in the mined results MD_K, this should not be known and visible to the external parties.
There are two ways to execute the Mine-and-anonymize approach:

○ One-step Mine-and-Anonymize

In this method, the data-mining techniques are required to be redesigned so to guarantee that the mined outcomes MD_K shouldn't violating k-anonymity for the original private table PT. This method wants to reshape data mining tools and techniques to implement k-anonymity directly and by integrating the two steps, it results in a more proficient process giving then execution advantages. The figure shows a graphical representation of one step mine-and- anonymize approach.

Figure 2: One-step Mine-and-Anonymize Approach

○ Two-step Mine-and-Anonymize

For the data-mining process, this approach doesn't want any change, so it can use any data mining technique which is available. In this approach, the result is required to be sanitized, removing such data from MD which compromises k-anonymity for the original PT. The Figure shows the graphical representation of the two-step Mine-and-Anonymization approach.

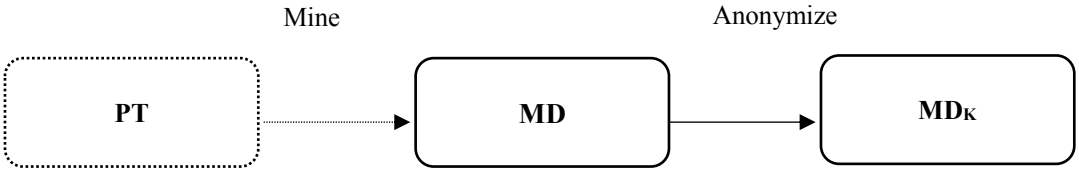


Figure 3: Two-step Mine-and-Anonymize

The main advantage of the mine-and-anonymize approach is the effectiveness of the data mining execution and the excellence of the results.

The leading disadvantage of this approach is it only allows data holders or the parties that are allowed to access the original private table PT to perform the mining process which may affect applicability.

- Attacks on k-anonymity

K-anonymity technique [15] is a well-known anonymization technique because it is simple and easy to anonymize data. However, k-anonymity is vulnerable to various attacks. If sensitive values in equivalence classes are not diverse, and the adversary has some background information/knowledge then k-anonymity doesn't provide privacy. Some well-known attacks on k-anonymity are as follow:

  - Homogeneity Attack

Inhomogeneity attack, for a delicate attribute, all values in a bunch of k-records are similar. Although, the data is k-anonymized still the adversary can predict the value of the delicate attribute for that bunch of k-records.

○ Background Knowledge Attack

In a background knowledge attack, to focus on probable values of the sensitive fields further, the attacker can use a connotation amongst quasi-identifier(s) with the sensitive attribute.

K-anonymity can protect against identity disclosure but it may not provide security against attribute disclosure. Therefore, the l-diversity technique was proposed which is capable of maintaining a minimum group size of k but also capable of maintaining the diversity of the sensitive attribute. The l-diversity privacy model of l-diversity is as follow:

• The l-diversity model

DEFINITION (l-diversity principle): "Let a $q*$-block be a set of tuples such that its non-sensitive values generalize to $q*$. A $q*$-block is l-diverse if it contains l well-represented values for the sensitive attribute S. A table is l-diverse if every $q*$-block in it is l-diverse" [17].

L-diversity is a privacy method which is protecting against attribute disclosure.

Possible attacks on l-diversity are:

○ Skewness Attack

This attack can occur when the adversary can derive sensitive information based on distribution of frequency of a sensitive attribute. I.e. in a released table RT, 1% of the record value of the sensitive attribute is HIV Positive, whereas
99% values of the sensitive attribute are HIV negative.

"Equivalence class 1: 49 positive + 1 negative"
"Equivalence class 2: 1 positive + 49 negative"

○ Similarity Attack

An Adversary can obtain significant information when in a uniform class the values of the sensitive attribute are different from each other but semantically similar in a particular quasi-identifier.

In the case of multiple sensitive attributes, for the curse of dimensionality, l-diversity becomes challenging especially.

Table 4: Original table (salary & disease)

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	Gastric ulcer
2	47602	22	4K	Gastritis
3	47678	27	5K	Stomach cancer
4	47905	43	6K	Gastritis
5	47909	52	11K	Flu
6	47906	47	8K	Bronchitis
7	47605	30	7K	Bronchitis
8	47673	36	9K	Pneumonia
9	47607	32	10K	Stomach cancer

Table 5: 3-Diverse version of original table (salary & disease)

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	Gastric ulcer
2	476**	2*	4K	Gastritis
3	476**	2*	5K	Stomach cancer
4	4790*	≥ 40	6K	Gastritis
5	4790*	≥ 40	11K	Flu
6	4790*	≥ 40	8K	Bronchitis
7	476**	3*	7K	Bronchitis
8	476**	3*	9K	Pneumonia
9	476**	3*	10K	Stomach cancer

- ## The t-closeness Model

DEFINITION: "(The t-closeness Principle :) An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness" [18].

The t-closeness model is a supplementary augmentation of the l-diversity model. The Earth Mover Distance (EMD) is used to find the distance between the two probability distributions. Within the case of numeric attributes, the t-closeness model shows to be more productive than the other privacy-preserving data mining techniques.

- ## Generalization property

If X and Y are generalized on table 'T1', supposing that X is more general than Y, furthermore, T1 satisfies the supposed criteria of the t-closeness model via Y, so therefore T1 also satisfies the supposed criteria via X.

- ## Subset Property

Let suppose Z is a set of attributes in table T1. If T1 satisfies the supposed criteria for t-closeness in correlation with Z, then T1 satisfies the supposed criteria for t-closeness in correlation with any set of attributes W in such a way that W is a subset of Z. To authorize t-closeness property, there is no such standard strategy. Each attribute is generalized individualistically. For sensitive attributes, Different protection levels cannot be specified. In the case of numerical sensitive attributes, the Attribute linkage can't be stopped. Degrading the utility of the data is the main drawback of this technique. k-anonymity [19], l-diversity [20], and t-closeness are well-known privacy techniques used for privacy-preserving data mining (PPDM). Hospitals [21] [22], banks [23], and many other organizations [24] use these privacy techniques to assure data privacy and data quality.

Table 6 shows the comparative analysis of the above-discussed techniques which makes the basic foundations of the research work [18] [17] [15]:

Table 6: Privacy Preserving Techniques are in brief

Privacy Preserving Techniques	Findings	Limitations
K-Anonymity	It face the conflict between information loss and disclosure risk. It prevents against the identity disclosure.	It may fail to protect against attribute disclosure.
L-Diversity	L-Diversity solve the problem of attribute disclosure.	It does not resist against the "probabilistic inference" attacks.
T-Closeness	It extends the l-diversity model by treating the values of an attribute distinctly by taking into account the distribution of data values for that attributes.	There is no computational procedure to enforce t-closeness property.

2.1.3 Distributed Privacy-Preserving Data Mining

The foremost objective of mostly distributed privacy-preserving data-mining techniques is to permit computation of valuable aggregate statistics of the whole data record without no compromising the individual's privacy. Therefore, the data contributors may work in collaboration to achieve aggregate results. However, they may not trust each other completely to share their private data. To solve the problem, the data sets may be partitioned horizontally or vertically.

2.1.3.1 Basic Cryptographic Techniques for PPDM

A distributed privacy-preserving data-mining problem has a close relation with cryptographic techniques which are used for generating secure results between different parties or data sharers. To achieve such a purpose, mechanisms of the "Secure Multiparty Computation (SMC) [25] area are applied. This example explains this situation well: two tycoons want to know that who is better off, without revealing their total income. This situation is about looking at data (numbers, digits, or alphabets) without disclosing it to others. Briefly, SMC is a method in which, multiple parties having their private data, desire to work together in such a way that their private information is secured and they get accurate results. I.e., a set of rules for conducting secure

elections, Security must be well-preserved to avoid adversarial conduct of the participants or an external party. This protocol is subjected to two types of adversarial models [25]:

- Semi-Honest

This type of adversaries is usually following the protocol honestly but they are curious and their curiosity led them to try to obtain the personal information of others from the data they received through the protocol. This is because they are called semi-honest adversaries. This model may be considered a realistic model of adversarial behavior in various conditions.

- Malicious

This type of adversary does anything to get the private information of other parties. They usually end ongoing protocol prematurely at any stage, send false and spoof messages to other parties to get their private information. Conspire with other malicious parties to breach the privacy of others.

The main problem with privacy-preserving data mining techniques which are currently used is that all participating parties should be honest, semi-honest, or malicious. Parties other than that cannot participate in the protocol. But there are many real-life conditions where the party's taking part in the protocols is “rational”. In other words, the parties will share their data-sets to obtain advantages and if required, they will act maliciously to increase their results. Such malicious actions will possibly affect the privacy-preserving distributed data mining protocols results.

The **semi-honest** models might be questionable for preserving the privacy of the participants in the protocol. The question is: if a party is trustworthy to participate in the protocol, then why it is not trustworthy for the data, this is helpful for the best outcomes of the technology. For example: for credit card fraud recognition, credit card corporations make data mining paradigms jointly. In such practice, the participating parties negotiate to view each other’s data. It’s not just about saving the data but also making the data secure is very important. It depends on the participating parties that they make it possible that

without viewing the data shared by other parties, this is how they can make the protocol secure. Similarly, the simplicity and efficiency that semi-honest protocols provide will help trusted parties to increase the security of their data but also secure data shared by other parties.

Informally, a trusted third party performing the computation is more secure and satisfies the principles of privacy, and it's the highly secured mutual computations. Suppose that every participating party shares their data with a trusted third party, and the trusted third party will work with complete secrecy and isolation and calculate the results without revealing the identity of individuals. After calculating and revealing the results to the participating parties, the trusted party will forget all the data it has seen. Not a single participating party can get information from the secured third party about any other participating party. But then again, no matter how much make the computations secure, a bit of information may be revealed about the data sharers.

Any sort of information that possibly is obtained from any individual's specific data and the result might be revealed by the protocol. There can be two types of information disclosures; the information disclosed from the results generated irrespective of the method used to compute the results, and the information disclosed from a particular process of calculating the results. The second type of information discloser is demonstrably predictable. This type of information is not disclosed because of the process.

- ## Composition theorem

The Composition theorem is one of the very beneficial theorems for the semi-honest model. Let's suppose, 'x' can be decreased secretly to 'y', and 'y' is calculated secretly by a trusted third party in the protocol. Another protocol calculated 'x' secretly. It shows that a protocol will be secured if sub-protocol is secured. In other words, sub-protocols can be grouped to produce a new secure protocol. Homomorphic encryption technique [26] is used to make sub-protocols. This encryption technique is used for generating a key from the set of polynomial-time algorithms that generates a public key "cryptosystem p [G, E, D]". Key generating algorithm G generates a private key (sk) and public key (pk). Any party can encrypt the message using public-key PK but cannot decrypt the message because private-key PK is required for decrypting the message, only the owner of data has the private key

SK. Encryption is performed by algorithm E which accept plaintext ‘m’ as input, random-value ‘r’ and public-key PK and generate output as corresponding cipher-text “Epk [m,r]”. Decryption is performed by algorithm D which accepts cipher-text ‘c’ as input and a private-key SK and generates output as a plain text “Dsk [c]”. It’s compulsory that “Dsk [Epk [m, r]] =m”.

2.1.4 Privacy Preserving of Application Results

An attacker can use the results of applications to create important implications regarding the behavior of the primary data. Many techniques for privacy-preserving data-mining tend to maintain the privacy of the final results of applications, these techniques are association rule-hiding and query processing. In statistical databases, such problems are associated with disclosure control [26] [25]. However, advancements are data-mining techniques that provide progressively mature approaches for attackers to make implications regarding primary data behavior. In such a situation, where the commercial data needs to be shared but it also has sensitive information which needs to be preserved from implication, for target-marketing reasons the association rules may disclose that sensitive data. The key purpose of these privacy-preserving methods are to prevent the attackers from creating implications from the final results of data-mining applications.

Techniques used for privacy-preserving of application results are as follow:

2.1.4.1 Association Rule Hiding

In the last few years, enormous developments have been made in the field of privacy-preserving application results to effectually implement association rule mining. Association rules frequently encode significant target marketing information regarding businesses; i.e. information related to credit card fraud [27].

For association rule hiding two main methods are used:

- Distortion

In this method [28], the input values are customized to a different value for a given transaction is. As we know that usually, binary transactional data sets are used to the input value is flipped in that case.

- Blocking

In this method [29], input values are not transformed but remained unstrained. Therefore, detection of association rules is avoided by using unidentified input values.

The distortion and blocking methods both have various drawbacks. As a result of such methods, some non-sensitive rules might be gone with sensitive rules, and new ghost rules may be generated. These drawbacks are unwanted as data utility is reduced by them for a mining purpose [25].

2.1.4.2 Downgrading Classifier's Effectiveness

For the data owner classification results may be very sensitive, for that reason this is a main privacy-sensitive application for classification. Thus, the problem is to reduce the accuracy of the classification system by altering the data, whereas the utility of data is not reduced and can be used for other types of applications. There are many possible ways [30, 31] to implement this method to decrease the efficiency of the classifier in perspective classification rules and applications of the decision tree. In the perception of obstructing implication channels for classification at the same time as mining the overall utility achieved, the system of "parsimonious downgrading" [30] was proposed. Using these principles, the "Rational Down grader" [31] system was proposed.

The strategies for association rule hiding can also be generalized to rule-based classifiers. For rule-based classifiers, the association rule hiding techniques can be used. Because, rule-based classifiers frequently utilize association rule-mining approaches, thus the rules having class labels, as a result, are used for classification. These rules are sensitive rules for the classifier downgrading method, and all other rules having non-class attributes, as a result, are non-sensitive rules. For reconstruction methods for categorical datasets, an algorithm [32] has been proposed for preserving the privacy of classification rules.

2.1.4.3 Query Auditing and Inference Control

Several databases having sensitive data don't share their data with the public, but there is a possibility that they might have some public interfaces through which they allow aggregate queries. This technique has a risk that a clever adversary may surmise sensitive information regarding the data by posing a set of queries. This insinuation may completely disclose whole data, an attacker may conclude the same values of the attributes of the data. The Second possibility is partial discloser in which an attacker may confine a range of values for an attribute value but maybe this doesn't help him/her to predict the exact value of an attribute of the data set. Most work of query auditing is usually focused on full disclosure.

To decrease the possibility of detection of sensitive data two methodologies are proposed:

- Query Auditing

  In this method, from a set of queries, one or more queries are not answered. In order to preserve the sensitivity of primary data, some queries are not answered [33] [34] [35] [36]. Two types of query auditing are:

  - Online Query Auditing

    In this method, the database doesn't know the set of queries in advance. Approaches used online query auditing is SUM query. In SUM [36] [37] query privacy is preserved by via

    $$(2.k - (q + 1))/m$$

    restricting the sizes and pair-wise overlays of the permissible queries Let's assume, the size of the query is limited to be as a maximum 'k', and the number of collective elements in pair-wise sets of queries is as maximum as 'm'. Therefore, if q is the number of elements that the adversary previously recognizes because of having background knowledge, then the most quantity of queries allowed are [27] [38] [39]:

- Offline Query Auditing

In this method, the database does know the set of queries in advance. From an auditing perspective, offline query auditing has better optimization. A set of queries is given in offline query auditing which has been responded to honestly, and it must be verified if privacy has been broken or not. For verifying whether a database follows its revelation properties [36], and offline auditing structure was proposed. The main idea behind creating an audit manifestation is it states entries for sensitive tables. There are some other approaches that are used for query auditing as bucketization [40] and histogram [41] [42] based methodology.

- Query Inference Control

In this method, the original data or the query results are distorted to preserve the privacy of the original data. The Pseudo-random sketches approach is used for query inference control to preserve privacy [43] [40].

In this approach, data is used to make sketches that illustrate the answer to the queries [40] [44] asked by the user. This approach is efficient for privacy preservation and it is also effective for providing data utility. An approach [45] is proposed for adding queries like addition, or a comparative frequency of random sampling that uses random data samples to calculate the sum function. The Sum of queries can be altered [46]. Noise is required to add to the results of queries [35] [47].

2.1.5 Limitation of Privacy

- Curse of Dimensionality

The curse of dimensionality is naturally inhabited by many methods of privacy-preserving data mining on the condition that information is made public. For the method of k-anonymity [48], there is an impact of the growing behavior of dimensions. When a borderline amongst pseudo-identifiers and sensitive attributes gets unclear as long as the

adversary has plenty of background information then it becomes the main concern. For the methods of l-diversity, it is also an incentive to make more privacy attacks using background knowledge [49]. In the case of preserving privacy, a bigger quantity of attributes are required to be suppressed which loses the data utility in the case of integrating data mining algorithms. The supposition at the back of the result is that extensive ranges are been made for the sake of generalizing attributes [48]. After generalization, attributes are sparsely inhabited in such a way that even 2-anonymity gets unlike immensely.

The l-diversity method becomes impracticable immensely to apply superbly with the increase in data dimensions.

2.2 Ensemble Learning

Ensemble learning is an area of research in which multiple learning machines are combined to solve a particular computational problem.[50] This technique is mostly used to improve the performance of overall system. In the recent years, ensemble learning has become one of the main research areas in pattern recognition and machine learning research [51].

2.2.1 Ensemble of Classifiers

There is no single classifier to achieve 100% accurate classification rules for all situations, due to this an ensemble of classifiers are used to achieve better performance. The Gathering of classifiers is a set of distinctive learning classifiers whose personal predictions are combined in such a way that it classifies modern illustrations to improve classification results. An ensemble of different learning classifiers mostly results in more effective classifications than a single classifier. Despite this, the question has often been discussed about which classifiers should be selected for a specified problem to produce the best possible ensemble.

Example:

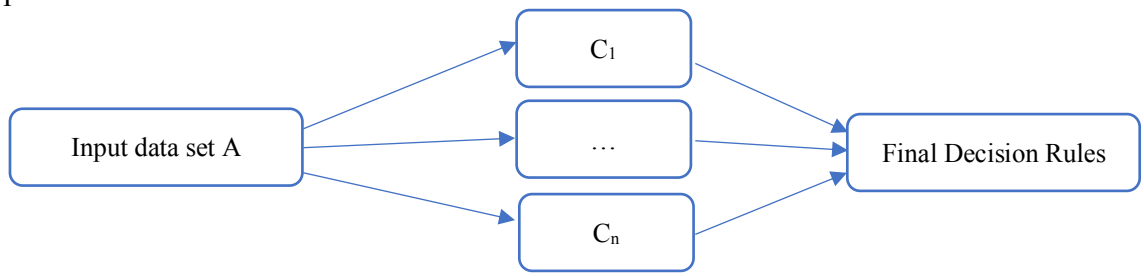


Figure 4: Pictorial Representation of an Ensemble Classifiers

Ensemble of classifiers is a popular area of machine learning research and a large number of experiments are reported. It consists of both, the combination of distinctive classifiers for the same feature set and the combination of classifiers for diverse feature sets. Several fixed and trained combining rules are used for the ensemble of classifiers. Ensemble of classifiers is a popular technique because bad classifiers and also bad feature sets may have useful information which can be used for improving performance by combining rules. By combining both, different feature sets and different classifiers the best performance can be attained.

The three main groups of ensembles of learning classifiers are as follows [52] [53]:

- ### Parallel combination of classifiers

  In this approach, classifiers are combined for data sets having different feature sets. This approach can be particularly beneficial for the objects that are represented by different sets of features, are portrayed in different physical domains such as visual and a wave of sound, or different sorts of analysis such as movements and frequencies. The original feature set can also be subdivided into subsets to decrease the dimensionality and the accuracy of a particular classifier. The parallel classifiers are often of a similar kind, but it is not necessary, they can be of a different kind.

- ### Stacked Combination of Classifiers

  In this approach, for the same feature sets, different classifiers are calculated. The nature of the classifiers may be different in stacked classifiers, for example; the fusion of neural networks, parametric decision rule, a closer proximity filer, etc.

- ### Combination of weak Classifiers

  In this approach, large groups of simple classifiers are trained on the modified versions of original data sets, for example; nearest mean rule and decision trees classifiers. The Two main approaches of this method are bootstrap bagging and boosting.

○ Bagging

This approach is also known as bootstrapping. In this approach [54], initially, an ensemble is formed by creating replicas of the bootstrap of the learning classifiers. Then, for getting an aggregate predictor, multiple hypotheses are engendered. By mediating the outputs in regression or by majority or weighted voting in classification problems, aggregation can be performed. Every classifier in the ensemble can vote with equal weight. The samples in bagging are obtained by replacement by using a uniform probability of distribution. In order to promote model variance, it directs every model in the ensemble with a subset of the training set which is haggard randomly.

○ Boosting

This approach gives the utmost importance to the most often misclassified examples of the previous basic learning classifier. By doing this, the base learning classifier focuses its attention on the most difficult examples. In boosting approach, it merges the base rules by picking the base rule having a majority weighted vote among the basic rules.

Experiments show that bagging is more efficient for noisy data than boosting approach.

2.3 Background Study of Genetic Algorithms

Genetic Algorithm is one of the well-known and recognized algorithms of EA [1]. GA is a population-based "Metaheuristic search technique" which is used to resolve optimization problems. Genetic Algorithms are developed to describe and replicate the mechanisms of natural selection and genetics that have been developed by Holland in 1975 [2]. A certain kind of problem enters on an optimization problem where a roughly and less time-consuming solution is admissible instead of a more precise but more costly[59]. Genetic algorithms represent an intelligent development of a random search system use up to solve optimization problems. The Genetic Algorithms are inspired by Darwin's evolution theory "the survival of the fittest", it's a general aspect of nature that in a contest, the "Fittest" individuals have domination over the weaker ones. In the present era of evolutionary computing, genetic algorithms hold one of the

important parts. Genetic algorithms represent an intelligent structure amongst the random search techniques which are implemented to solve the optimization problems, which are simple and easy to execute. For any specific problem, a genetic algorithm works for finding a solution by simulating natural processes, like selection, crossover, and mutation, to develop an effective and optimized solution for a given problem [55].

2.3.1 Search Space

A population of individuals has preserved in search space for a genetic algorithm, every individual in a population signifying a potential solution to a given problem. Every individual is entitled as an element (generally a binary number, an integer, or a real number) of a fixed-length vector generally it depends on a given problem and population type. To persist the genetic likeness, these individuals are associated with a chromosome and the variables are associated with genes. In this way, a chromosome (solution) is made of some genes (variables). A fitness function is used to evaluate every chromosome indicative of the aptitudes of an individual to "compete". The individual having the optimal (or generally a near-optimal) fitness criteria is seek out. An objective of using a genetic algorithm is selective "breeding" of the solutions by combining information from the chromosomes to generate "offspring" which is fitter than the parents.

A genetic algorithm has a population of n chromosomes (solutions for a given problem) and a fitness function that contains the related fitness criteria. For reproduction, the fittest chromosomes (parents) are selected from the population on the basis of their fitness for producing offspring through a reproductive strategy. As a result, fittest solutions are providing furthermore chances to reproduce, with the purpose that offspring inherit the characteristics of each of their parent. Space must be made for the new offspring since the population is kept at a static size and parents produce new offspring, they need space in a population. Weak or less fit individuals in the population die, and they are replaced with new solutions (individuals). The Algorithm creates a new generation when all mating opportunities in the old population are consumed. It is hoped that over a number of successive generations a better solution will succeed and the least fit solutions die-off.

Every successive generation contains the finest "partial solutions" as compared to the earlier generations. Once the population has converged and is producing offspring containing solutions

the same as the previous generation or repetition of results occurs, again and again, the algorithm is said to have converged to a set of solutions to the given problem.

2.3.2 Operators of GA

A genetic algorithm uses genetic operators for preserving genetic diversity. It is very significant to preserving genetic diversity for the procedure of producing the best solution for a given problem. The Genetic operators of GAs are the same as natural process because GAs is inspired by genetic structure of nature.

A genetic algorithm consists of three types of operators which are selection, crossover, and mutation. Genetic operators of GAs are as following:

- Selection

How to select a parent solution? A GA tends to select better, stronger, fittest solutions for breeding, as happens in Darwinian evolution.[60] The Selection represents the principle of "survival of the fittest". The Fittest chromosomes survive and are selected for reproduction. Weak chromosomes have a low probability to be selected for reproduction. Selection is basically taking out the subset of genes from a present population is based on some fitness function or fitness criteria. The fitness function is fitness which measures the fitness of an individual to be selected as a best result or solution for a given problem, or select the fittest chromosomes for next generation.

- Crossover

Crossover/recombination is a genetic operator, in which two parents (chromosomes) are selected from the population by "the selection operator for producing a new offspring (chromosome). A crossover operator sets a crossing point along with the bit strings is randomly chosen then the values of the two strings (parents) are swapped up to this location. The two new offspring are produced from this process, new offspring are added to a population for the later generations. Common methods of crossover operator which are uses for selecting parents to perform crossover are:

- Tournament selection
- Roulette wheel selection

- Rank selection
- Steady-state selection.

The basic conception of crossover is, after reproducing every parent chromosome that is chosen based on a fitness function, the offspring chromosomes will be fitter because they inherit the best characteristics of their parents. The crossover is likely to create even better individuals as a result of recombining sections of good individuals.

Example:

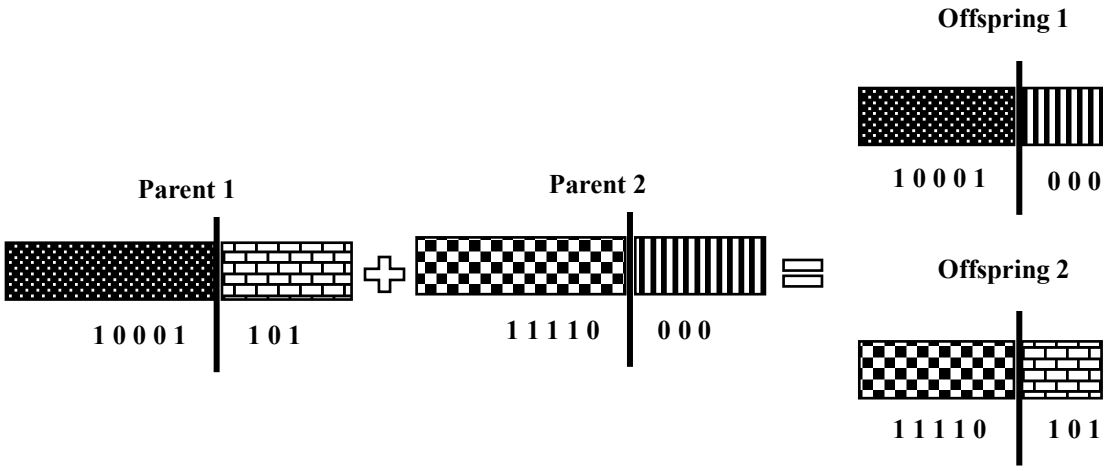


Figure 5: Example of Crossover

Crossover has many types like the single-point crossover, two-point crossover, uniform crossover, and arithmetic crossover.

• Mutation

Mutation brings in random adjustments. A mutation occurs in the evolution phase where the user specifies the probability of mutation. The Mutation is about exploring new areas in the search space, which has the effect of avoiding convergence.[61] The probability of mutation is generally set to a relatively small value as 0.01 which is a decent first pick. A transformation is a genetic operator, and its objective is to sustain genetic diversity from

the one generation of a population of chromosomes (individuals) to the next generation and inhibit premature convergence.

Example:

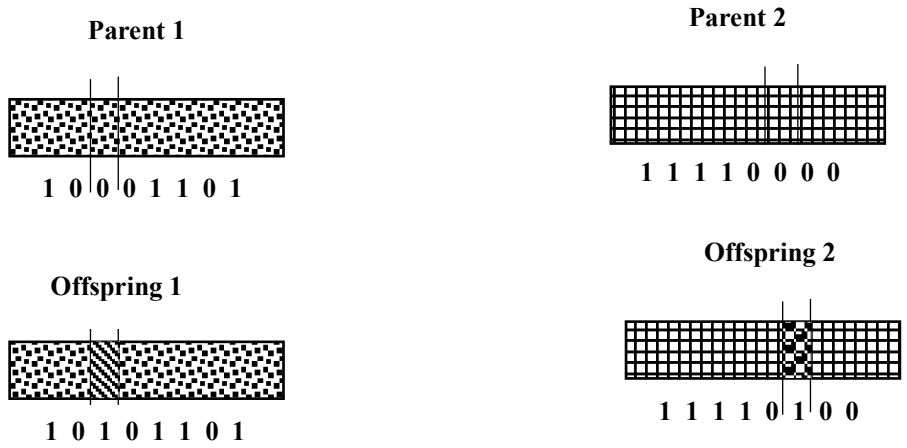


Figure 6: Example of Mutation

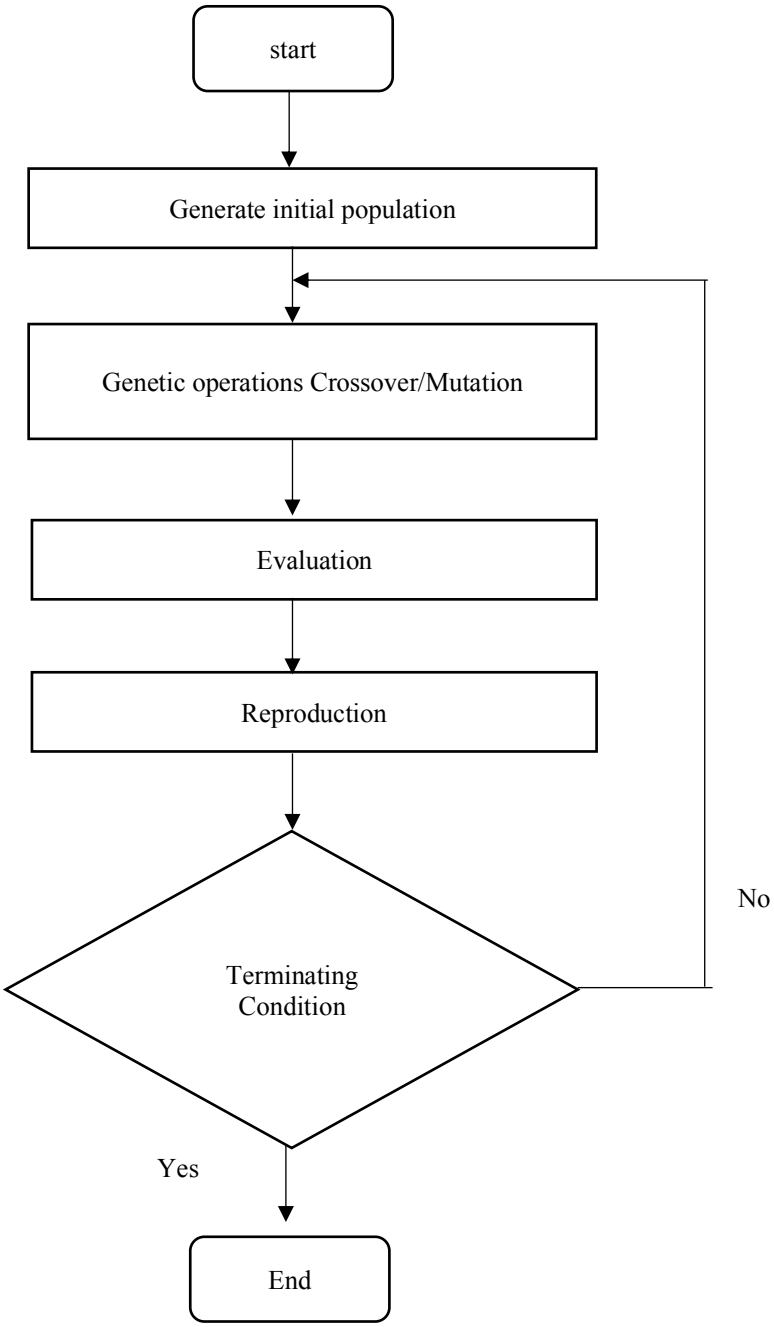


Figure 7: Flow chart of General Genetic Algorithm

Chapter 3

Proposed Technique and Implementation Details

3 Proposed Technique and Implementation Details

This system is proposed in arrange to address the concerns related to protection and privacy. Regarding that; taking an interested party to make an understanding on characterizing the level of sensitivity of attributes and taxonomy for categorical data types. It needs to pass through different steps to make it useful as well as private. The released table RT containing attributes A1, A2,…,An of categorical and numeric types passes through different phases of transformations. These transformations are made during the classification process, where rules are generated and instances are generalized simultaneously. In a distributed environment, different factors add more concerns to the process of maintaining the integrity and efficiency of classification rules. Network protocols require the support of different cryptographic techniques to maintain privacy and authenticity among multiple sites. In this research work, cryptographic techniques are not taken into account in the development of privacy framework but it is developed by keeping in mind, that its interactions with the cryptographic layer is possible by the addition of few components. The major focus was to generate accurate classification rules and making them private to prevent privacy attacks at the same time.

In the varied area of privacy-preserving data mining (PPDM), researchers are working on making data private using different privacy techniques like partitioning data, generalizations and data distortions, and then forward it to the mining process for data exploration and analysis. Some researchers are oppositely; they perform the mining process at the data first, and after that transform it with different privacy-preserving techniques. Both methods require a trade-off between privacy and accuracy. The proposed approach is placed located in between these two research areas where privacy-preserving techniques and classification tend to generate the final results in the form of anonymized rules and instances while preserving the correspondences of sensitive attributes and key attributes.

In this chapter, each process of the proposed technique is described in a more elaborative way including their computational procedure. In this domain, different research papers test the privacy properties between the sensitive and key attributes using different distance measurements relating to distributions, but no appropriate computational method is given. If it is given, it would greatly damage the usefulness of the data.

This framework consists of different processes from generating classification rules from multiple classifiers like j48, Ripper, and Part based learning algorithms [5]. Calculating distances between rules. Optimizing the threshold value of t-closeness with a Genetic Algorithm. Group similar classification rules using clustering techniques based on the distances of the classification rules. Once rules are grouped, instances are mapped to the rules accordingly. In the last phase, sensitive instances and rules are filtered out separately from the buckets to implement the generalization process according to the selected group anonymization technique.

Following are the detailed information of step-by-step execution and transformation processes of the proposed technique:

3.1 Problem Formulation

Following are the definitions used in the framework:

3.1.1 Definition (Released Table RT)

Release Table (RT) is a training data set available for classification problems. It can be taken from UML Repository [56] or provided by various data sources. There are many ways to describe the released table. Usually, tables are used for RT. A Table is a collection of records and every row contains complete information of one single record or instance, it also may have useless instances. Each record in a table is characterized by a sequence of attributes such as explicit attributes, quasi-identifiers, sensitive attributes, and target class.

Let's represent the Training data set released table as RT, having N number of attributes. RT= (a1, a2,….,an) and class attribute represents the target class. There can be two types of attributes, categorical attributes, and numerical attributes.

3.1.2 Definition (Quasi Identifiers QI)

Quasi-Identifiers denoted as QI is a set of attributes. QI can be linked with external information so that few or whole respondents are re-identified to where the information points out. QI is a subset of attributes of RT.

3.1.3 Definition (T-Closeness)

"An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t- closeness if all equivalence classes have t-closeness." [18]

3.1.4 Definition (Taxonomy of Attributes)

There exists a natural notion of measuring distance for numerical attributes, but categorical attributes do have not a clear method of calculating distances. As a solution for this problem, the domain experts built the hierarchy of generalizations for the attribute. With Taxonomy anonymization of the records, it's easy by moving through a different level of generalizations in the hierarchy. Descendants are considered when the attribute value is required to replace with a more specific value. Ancestors are considered when the attribute value is required to replace with a more generalized value.

3.1.5 Definition (Equivalence Class)

An equivalence class is a set of similar anonymized data instances.

3.2 Methods of the Framework

Following are the processes of the framework:

3.2.1 Process 1: Rules Generation using Multiple Learning Classifier

The selection of learning classifiers is based on the nature of supplied data set as input. For the developed framework, multiple classifiers are used to generate classification rules. Here a question arises, why use multiple classifiers? No single classifier produces 100% classification rules. That's the reason why multiple classifiers are used in this framework to generate rules. When rules generated from different learning classifiers are combined it gives a strong classification result that cannot be obtained from single classifiers.

In this framework classification rules are generated from Ripper, J48, and PART classifiers.

- Ripper

The "Repeated Incremental Pruning to Produce Error Reduction (RIPPER)" algorithm was introduced by Cohen in 1995 [3]. This algorithm is an optimized form of the IREP algorithm. Ripper also known as Jrip is a well-known and reputable algorithm used for the classification of supervised data. It works very efficiently on noisy data. Rules generated from ripper are easy to understand. The technique of ripper classification is simple. Firstly, the training data is randomly distributed into two sets called growing set and pruning set. After that, each rule continues growing until there is no more information gain is possible.

```
Procedure IREP(Pos,Neg)
begin
    Ruleset := θ
    while Pos ≠ θ do
    /* grow and prune a new rule */
    split (Pos, Neg) into (GrowPos, GrowNeg)
        and (PrunePos,PruneNeg)
        Ruleset := GrowRule(Rule,PrunePos,PruneNeg)
        Ruleset := PruneRule(Rule,PrunePos,PruneNeg)
        if the error rate of Rule on
            (PrunePos,PruneNeg) exceeds 50% then
                return Rulset
            else
                add Rule to Ruleset
                remove examples covered by Rule
                from (Pos, Neg)
        endif
    endwhile
    return Rulset
end
```

Figure 8: Ripper Algorithm [3]

- PART

Global optimization is performed in Projective Adaptive Resonance Theory (PART) and good rules can be learned one rule at a time. The basic working of PART is like it constantly generates partial trees and finds rules [4]. PART is consists of two foremost theories, one is "rule generation", e.g. creating rules using decision trees, and the other is the "separate-and-conquer" rule-learning method. PART is a simple and efficient algorithm, and also this avoids post-processing that's the reason it is more appreciated.

- J48

J48 (c4.5) is a tree data structure classification algorithm, in which leaf nodes represent target classes and internal nodes represent test conditions. It constructs trees in a top to bottom manner using a greedy algorithm. The purpose of this algorithm is to fits the training data by constructing a tree. The tree starts initially with a single node. Information gain is used as a heuristic in this algorithm to selecting those attributes which are best for separating, which split up the training samples according to their classes. Information gain helps a tree to choose an attribute that is determined to have the highest access to information. The selected attribute turns out to be a test case or a branch of the tree. To every attribute value, a branch (node) is created in a tree. Attribute once selected, not considered again at any descendants of the node. Once a tree is completed, each pathway from the root node to the leaf node comes to be a rule. The leaf nodes of the tree correspond to a class label of the rule [5]. J48 (c4.5) algorithm can handle both continuous and nominal attributes. In the case of the continuous attribute, it creates a threshold and then divides the training samples to those who are above greater than the threshold and those having a value less than or equal to the value of the attribute. This algorithm can also deal with missing values and attributes including different costs. Once a tree is constructed tree pruning is performed.

1 Check for base cases

2 For each attribute α

    a. Find the normalized information gain from splitting on α

3 Let α_best be the attribute with the highest normalized information gain

4 Create a decision *node* that splits on α_best

5 Recurs on the sub lists obtained by splitting on α_best, and add those nodes as children of *node*

Figure 9: J48 Algorithm [5]

3.2.2 Process 2: Combining Classification Rules

Once classification rules are generated by WEKA, those rules need to be converted from WEKA generated output to an excel document for further processing. As discussed earlier, three classifiers are used to generate rules, so for every dataset, three separate rules files will be generated. Initially, every rule file will be converted separately from the WEKA file to the excel file. In the next step, those separate rules files will be combined into a single excel document.

3.2.3 Process 3: Find Distance Between Rules

Once the rules are converted from the WEKA file to an excel document, the next step is to calculate the distance between rules. In this step, the difference will be calculated between every rule from the other rule. To hold the t-closeness property, we need to calculate distance between attributes within the equivalence class and in the entire table. To put it simply, according to the closeness property similar attributes are grouped in a single place. In the proposed technique, similar rules are grouped together by applying this concept. Every single rule is compared to the other rules in the rule file.

The following algorithm calculates the distances between rules:

```
ans = 0
for i in range (0, size[0]):
    for j in range (i+1,size[0]):
        ans = 0
        for k in range (0,size[1]):
            if (Data.iloc[i,k] = = Data.iloc[j,k]):
                ans = ans + 0
            else:
                ans = ans + 1
        d[j][i] = ans/size[1]
        d[j][i]
```

Figure 10: Code for calculating distance between rules

Distance calculation of two rules are as following.

Consider two rules R1<A1….An> and R2<A1….An>, having attributes A1 = <n1, v1> and A2 = <n2, v2> (n = attribute name, v = attribute value). A distance counter is used to calculate the dissimilarity between attributes, add 1 to the counter if dissimilarity is found, after comparing all attributes of the two rules divide the counter value by the total number of the attribute.

Example: To calculate the distance of two rules, their attribute values will be checked. Initially set distance counter value to zero. Supposed that two attributes having the same name will have the same type of values (numeric or categorical). The distance D (R1, R2) will be calculated as follows. Initially, compare the first attributes of both rules, if R1 (n1) ≠ R2 (n1), i.e., the attributes have different names, then compare the attribute of rule 1 with the next attribute of the second rule. If R1 (n1) = R2 (n1), then compare the attribute values. Attribute values are considered as a string here so the string matching technique is applied here. If dissimilarity is found, add 1 to the distance counter.

$$D = \frac{\square\square\square\square\square\square\ \square\square.\square\square}{\square i\square\square i\square\square\square\square\square\square\square\square\square\square\square\square\square i\square\square\square\square\square\square}$$

Table 7: Example of Distance calculation of rules

	Age	Gender	Zip Code	Disease
Rule 1 (R1)	45	Female	44000	Diabetes
Rule 2 (R2)	34	Female	44000	Pneumonia

Distance of rules of (Table 7) is: 1 + 0 + 0 + 1 = 2/4 =0.5. The Total number of attributes are 4, values of age and disease attribute are different so 1 is added to the distance counter for both attributes. Whereas, values of gender and zip code are the same so zero is added to the counter for these attributes.

All the Rule(s) are compared using the above-mentioned distance calculation method.

3.2.4 Process 4: Optimizing t-closeness with Genetic Algorithm (GA)

GA is used in the proposed technique to optimize the threshold value of t-closeness. It finds every possible threshold t and selects the best optimal t according to provided data type.

Example:

(A)

```
Output: matrix( [ 5,  7,  11 ], [ 7,  9,  10 ], [ 3,  4,  10 ], [ 3,  6,  11 ], [ 3,  7,  10 ],
[ 3,  10,  11 ], [ 4,  6,  10 ], [ 5,  8,  11 ], [ 3,  4,  10 ], [ 5,  6,  11 ], [ 7,  8,  11 ],
[ 6,  7,  10 ], [ 7,  10,  11 ], [ 5,  9,  10 ], [ 5,  6,  7 ], [ 4,  6,  10 ], [ 3,  7,  8 ],
[ 6,  7,  10 ], [ 4,  7,  8 ], [ 3,  6,  10 ], [ 4,  7,  9 ], [ 6,  7,  8 ], [ 7,  8,  11 ],
[ 4,  5,  10 ], [ 6,  9,  11 ], [ 7,  9,  10 ], [ 4,  7,  10 ], [ 5,  8,  9 ])
```

(B)

```
In:     print fit.min()
        print ptt[best]

        0.111111111111
        [[  4  7  10  ]]
```

Figure 11: GA generates every possible t (A) and select the best optimal t for (Table 4)

GA is an evolutionary algorithm well-known for solving problems related to optimization. GA is selected for the proposed framework because of its global search method which is easy, simple, efficient, adaptive, and robust. Its capabilities as a machine learning technique are acknowledged in the information retrieval domain. GA is easy to adopt and its chromosome structure is very simple as compared to the other optimization algorithms.

- Working of GA for the Framework

The Population of GA is consists of attribute values, population size is 500, and the number of generations are 500 for the proposed framework. The Fitness function is to find the minimum threshold vale of t-closeness. Threshold value are generated through the "Earth Mover's Distance (EMD)" formula. EMD [57] is a distance formula that evaluates distance among two probability distributions over an area D.

- EMD for categorical attributes

$$hierarchical - dist(Vi, Vj) = \frac{level(Vi, Vj)}{H}$$

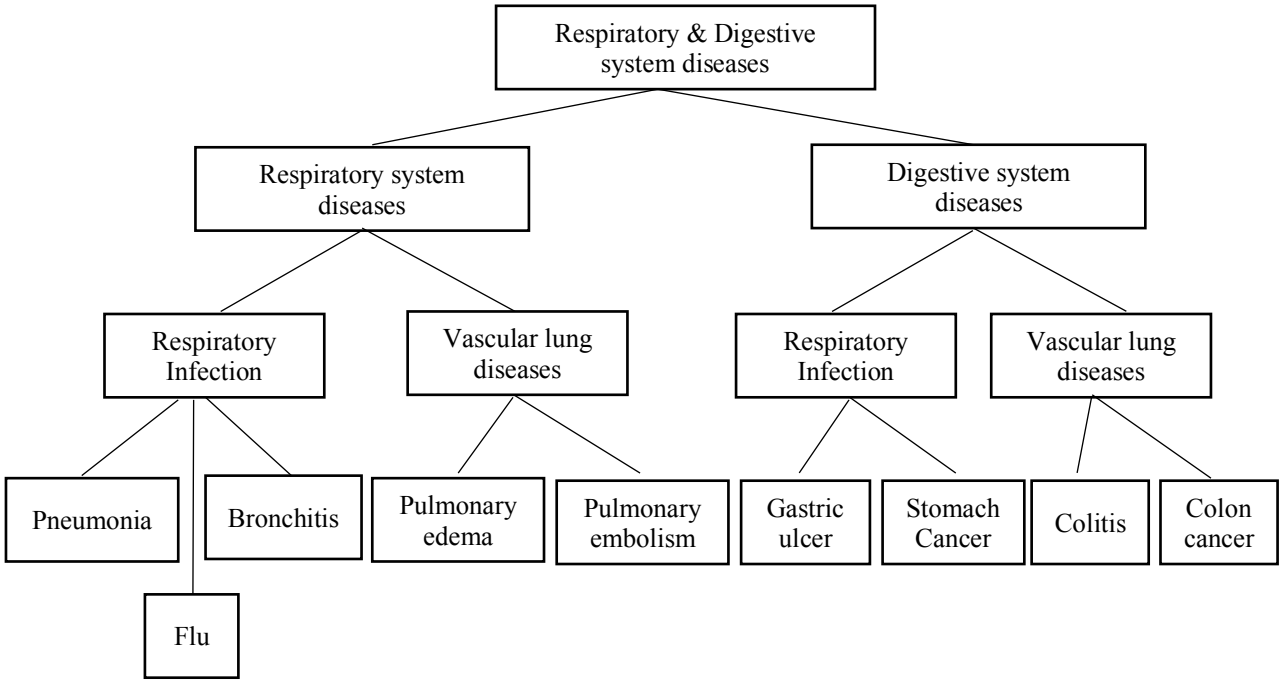


Figure 12: Hierarchy for categorical attributes Disease for (Table 4)

[**P**,**Q**] is calculated as:

$$D[P, Q] = \sum_{N} cos\ t(N)$$

- EMD for numerical attributes

Let $r_i=p_i-q_i$, then $D[P,Q]$ is calculated as:

$$D[\boldsymbol{P},\boldsymbol{Q}] = \frac{1}{m-1}\left(|r1| + |r1+r2| + \cdots + |r1+r2+\cdots+rm-1|\right) = \frac{1}{m-1}\sum_{i=1}^{m}|\sum_{j=1}^{i} rj|$$

This formula calculate distance between two distributions P and Q to optimize threshold value of t-closeness. Following example show

Example:

$$\boldsymbol{Q} = \{3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k\}, \boldsymbol{P}_1 = \{3k, 4k, 5k\}$$

Move 1/9 probability for each of the following pairs

$$(5k \to 11k), (5k \to 10k), (5k \to 9k), (4k \to 8k), (4k \to 7k), (4k \to 6k), (3k \to 5k), (3k \to 4k)$$

The cost of this is:

$$\frac{1}{9} \times \frac{6+5+4+4+3+2+2+1}{8} = \frac{27}{27} = \frac{3}{8} = 0.375$$

GA generates every possible combination of P and finds its distance from Q, the best P will be selected which satisfies the criteria of the fitness function and that will be the optimized threshold value of t-closeness. Uniform crossover is used to generate new offspring. The Probability of crossover is 0.8 and the probability of mutation is 0.2 for this framework. Tournament selection is used for selecting the population of the next generation. In this method, parents are compared with the newly generated offspring and the best solution is selected for the next generation. This algorithm runs 25 times to produce the best optimal value of t-closeness.

Following are the steps of the genetic algorithm for the proposed framework:

Step 1: Create an initial population of individuals randomly (first generation).

Step 2: Fitness evaluation of individuals using the EMD formula. Select the minimum values.

Step 3: As there is no proper function to measure the ideal t-closeness value so this step will be repeated for 500 runs to get the near possible solution. Stopping criteria is when the same

results appear repeatedly. Reiterate this generation until termination criteria matched (Adequate fitness is attained)

1. Reproduction: select the best fit individuals as parents for reproduction.
2. Crossover and mutation: Produce new offspring using crossover and mutation with 0.8 and 0.3 rates respectively.
3. Add new offspring to the population.
4. Go to step 2 for evaluation of new population

Step 4: output: the optimized value of t.

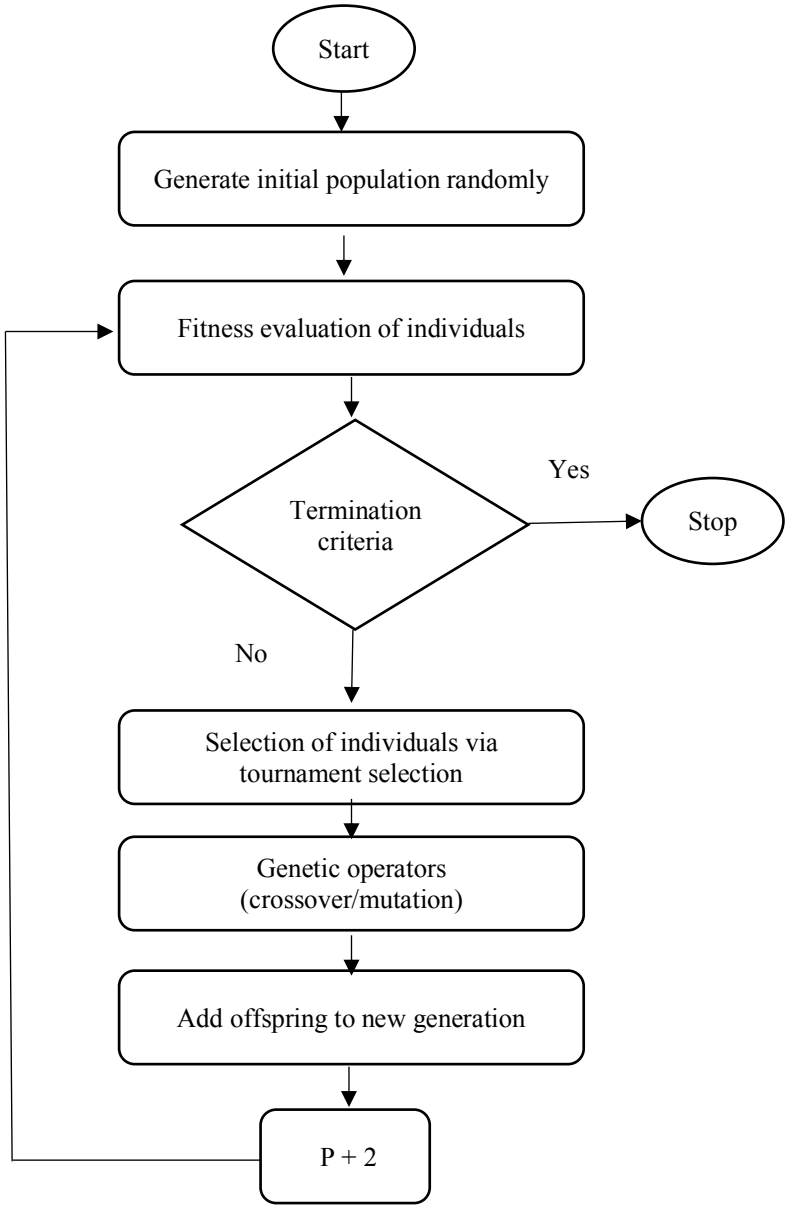


Figure 13: Genetic Algorithm Flow chart of proposed technique

3.2.5 Process 5: Clustering Rules using Distance Measures

After the distances are calculated, rules must be grouped in equivalence classes. These equivalence classes are classified using an optimized threshold value "t" along with range values [r1, r2]. Each rule is checked for its crisp membership with the appropriate class and added accordingly. These classes are labeled as "Rule Buckets" where range values are inclusive in the threshold compatible checking.

3.2.6 Process 6: Validation of Rules

Rules generated from the classification process need to be analyzed, rules are validated using the validation set, validation set contains instances that were separated in the initial phase. Now, weights are assigned to the rules in the rule file to evaluate the correctness of the classification system in conditions of accuracy and performance. The sum of correctly classified instances in the given set is called the confidence of the rule. The Confidence value of the rules are used to calculate the weights. The overall accuracy of the classification rules set is affected by incorrectly classified instances.

3.2.7 Process 7: Extraction of Sensitive Rules

After the rules and instances are grouped in the rule buckets, equivalence classes containing sensitive rules are filtered out separately. Sensitive rules or sensitive attributes having values, which the defendants do not want to be disclosed i.e. in the hospital table, patients having diseases that cause death are sensitive like cancer and HIV. Classes containing sensitive attributes are marked high-privacy zones, instances, and rules of these sensitive classes are needed to be more generalized than the other classes having less sensitive attributes and instances. Filtering out sensitive rules helps to preserves the level of privacy in different equivalence classes. As a result of this, depending upon the sensitivity of the rules and instances every class has a different level of privacy.

3.2.8 Process 8: Generalization of Rules

The rules and instances containing sensitive attributes are required to be generalized. Attributes (Quasi Identifiers) are considered for generalizations. To generalize the sensitive rules and

instances, having the details of the attributes is very important that whether a specific attribute is numeric or categorical. Categorical attributes are replaced by selected generalization possible for that attribute using a given taxonomy tree. Numeric attributes are replaced with a range of values.

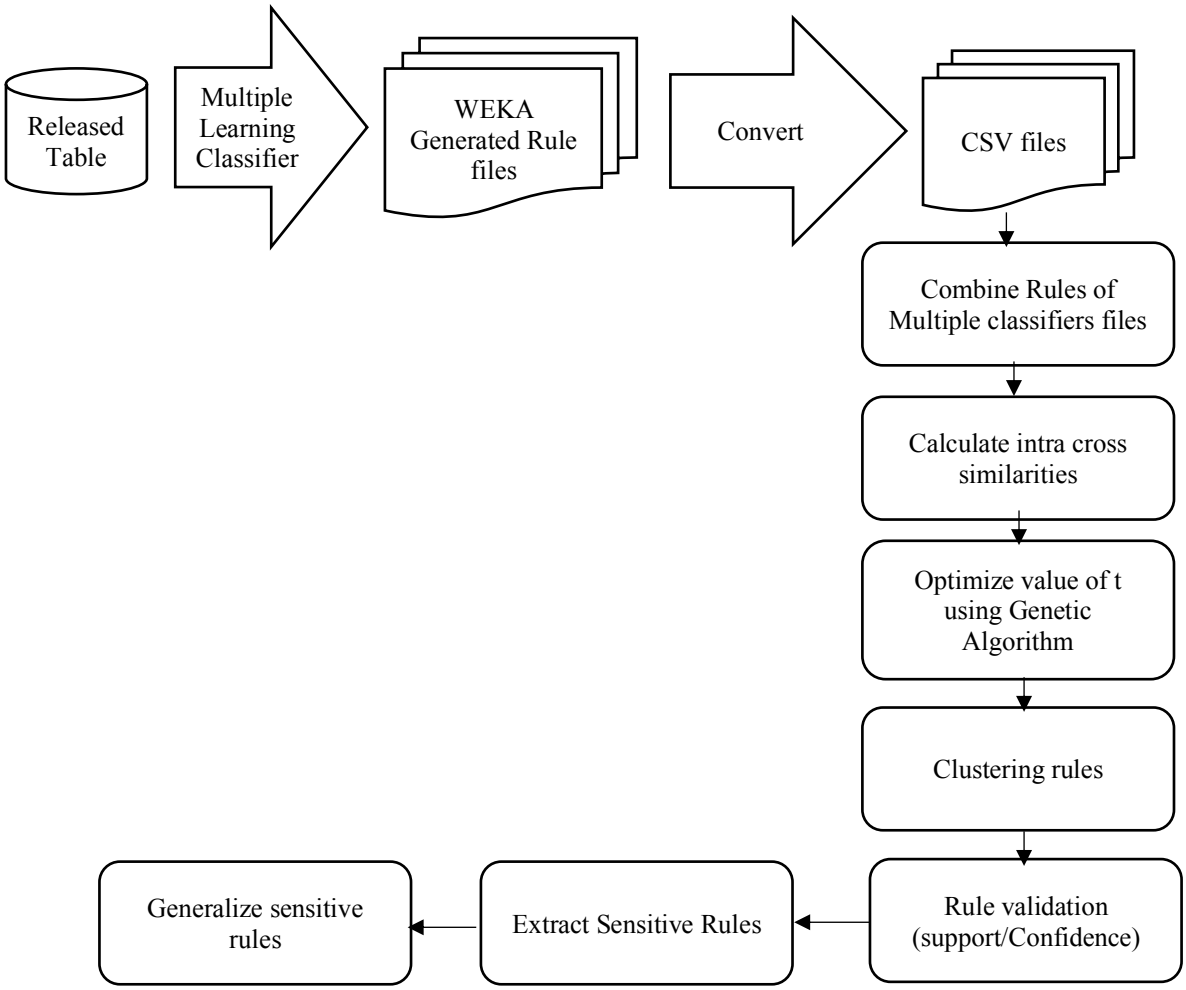


Figure 14: Architectural Diagram of proposed

Chapter 4

Experimental Results

4 Experimental Results

After completion of the implementation process of the proposed framework, the next step is testing. Testing the results generated from the implementation process is the most critical task to analyze the performance and accuracy of the research work. The Percentage split method is used for the statistical analysis. This method divides the raw dataset into three major sub partitions, marked as "Training set", "Validation set" and "Test set". "Training set" is used up for the classification process, in which it passed through the learning classifiers to produce the output. After generating rules, weights are assigned to rules on the basis of the validation set. Lastly, to analyze the performance and correctness of the generated rules, they are tested on the testing set. The testing set holds totally unseen data samples which are separated out in the early stage and never processed before. For the classification process, data must be in WEKA format [58]. Data sets used in the proposed framework are taken from UCI Repository [56].

Table 8: Data sets

Datasets
ADULT
Car
Heart Disease
Zoo
Balance Scale
Tic Tac Toe

Percentage splits of the raw data sets used for evaluation of this framework is shown in (Table 9):

Table 9: Percentage Split of datasets into Training, Validation, Test Sets

Training Set	Validation Set	Testing Set
66%	17%	17%

As discussed before, multiple classifiers are used for the classification process, decision trees (J-48), Ripper and PART are selected for the statistical analysis.

4.1 Experimentation Structure for Genetic Algorithm

A Genetic algorithm is used for optimizing the threshold value of t-closeness. The experimental structure for GA used in a proposed technique is shown in (Table 10).

For the proposed technique the defined population size is 500 for 100 generations and the defined size of the tour is 6. The size of the chromosome is an attribute value. Every possible combination of P is generated with GA and selected the P having minimum t value as output and the optimized threshold value of t-closeness. EMD formula is used for generating t values. Uniform crossover is used to produce new offspring and the probability of crossover is defined as 0.8. The Probability of mutation is defined as 0.2 and gene mutation probability is defined as 0.5 for the framework. Tournament selection is used for creating a new generation. In this method, new offspring and their parents are compared and the fittest chromosomes are selected for the new generation. Elitism is keeping best, fittest chromosomes are selected for the next generation. With different combinations of these parameters, experimentation has been performed. For the experimentation process, the finest values are preserved.

Table 10: Genetic Algorithm parameters and its values

Parameters	Values
Size of Population	500
Generation	100
Size of Chromosome	Attribute Values
Fitness Function	Min t closeness value
Selection	Tournament
Crossover	Uniform crossover
Probability of crossover	0.8
Probability of mutation	0.2
Probability of gene mutate	0.5
Elitism	Keep-best
Runs	25

4.2 Statistical Analysis Data sets

Following are the statistical analysis of data sets for the developed framework:

4.2.1 ADULT Data Set

The ADULT dataset is extricated from the census bureau database. This dataset contains about 32000 observations and having 15 attributes. The target class attribute predicts whether an "individual" has an income greater than or less than $50,000 a year. This database is used in previous t-closeness, and privacy-preserving techniques. The Attribute information of the adult dataset is given in (Table 11).

Table 11: ADULT Dataset attribute information

Attribute	Values
Age	17 – 90
Work class	Private, self-emp-not-inc, self-emp-in, federal-gov, local gov, state-gov, without-pay, never-worled
Education	Bachelors, some-college, 11^{th}, HS-grade, prof-school, assoc-acdm, assoc-voc, 9^{th}, 7^{th}, 12^{th} ,masters, 1^{st} -4^{th} , 10^{th}, doctorate, 5^{th}-6^{th}, preschool
Country	41 Countries
Marital Status	Married-civ-spouse, divorced, never-married, separated, widowed, married-spouse-army, married-AF-spouse
Race	White,black,Asian,-pac-islander,amer-indian-eskimo,other
Gender	Male, female
Salary	<= 50k, >50k
Occupation	Tech support, craft-repair, sales, exec-managerial, prof-specialty, handlers-cleaner, machine-op-inspct, adm-clerical, farming-fishing, transport-moving, priv-house-serv,others

Few attributes are tagged as privacy parameters. (Table 12) shows the assumption made to check the validity and correctness of the designed framework.

Table 12: ADULT Dataset Privacy Parameters

Privacy Parameters	Tagged Attributes
Quasi-Identifiers	Age, work-class, education, country, martial-status, age, gender
Sensitive Attributes	Occupation

After the percentage split, 66% of the training data set is passed to multiple learning classifiers and it generated 1143 classification rules collectively after the elimination of the rules with zero weights (no contribution to the classification system). Few rules which are classified as no class values are characterized as null.

(Table 13) shows the statistics of the final results of the framework:

Table 13: Statistics on Test Set of ADULT Data Set

Criteria	Results (Before Processing)	Results (After Processing)
Correctly Classified Instances	42229	48208
Incorrectly Classified Instances	1827 (3.7415%)	0
Unclassified Instances	4786 (9.8%)	634 (1.3%)
Total Number of Instances	48842	48842
Overall Performance (Accuracy)	86%	98.6%
Total Number of Generated Rules	1848	1143
Pruned Rules	0	705

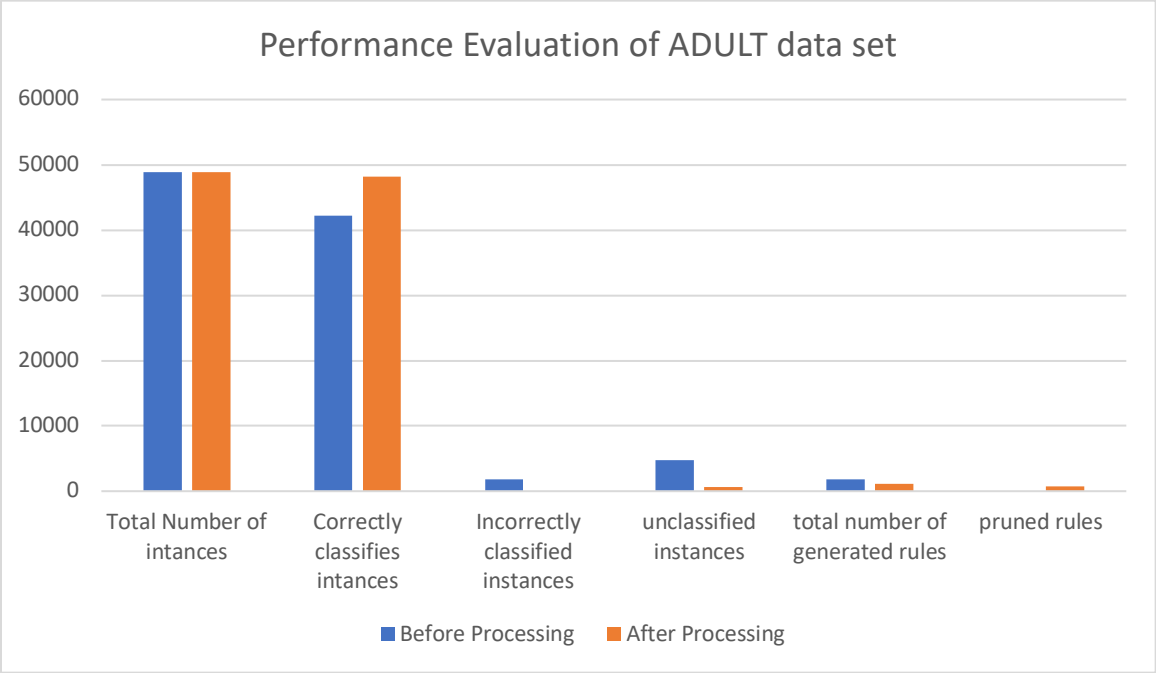


Figure 15: Performance evaluation of ADULT Data Set

(Figure 13) shows the performance evaluation of the adult data set.

During the classification process, the bucket size and range is decided. Classified rules are grouped in these buckets according to the threshold value. For the ADULT data set, a total of 7 buckets are produced along with the range difference of 1. The privacy results generated after the processing of the proposed work are as shown in (Table 14):

Table 14: Privacy statistics of ADULT Data set

Privacy Metrics	Results
Total no. of buckets	7
Total no. of sensitive rules	834
Privacy Preserved percentage	81.5
Maximum Similarity Measure	7
Minimum Similarity Measure	0
Bucket Range Difference	1
k-Anonymity	K=2

4.2.2 Car Data Set

This database is derived from a simple hierarchical decision model which assesses cars according to the specific structure which is given below. It doesn't contain the structural information and directly relates car information to the six attributes titled as: "buying", "maint", "doors", "persons", "lug boot", "safety". (Table 15) shows the attribute information of the car data set.

Table 15: Car Evaluation Dataset Attribute information

Attributes	Values
Target class	Unacc, acc, good, vgood
Buying	Vhigh, high, med, low
Maint	Vhigh, high, med, low
Doors	2, 3, 4, 5, more
Persons	2, 4, more
Lug Boot	Small, medium, big
Safety	Low, medium, high

Few attributes are tagged as privacy parameters. (Table 16) shows the assumption made to check the validity and correctness of the designed framework.

Table 16: Car Evaluation Dataset Privacy Parameters

Privacy Parameters	Tagged Attributes
Quasi-Identifiers	Buying, Maint
Sensitive Attributes	Maint, Target Class
Key Attributes	Persons (Suppressed)

After the percentage split, 66% of the training data set is passed to multiple learning classifiers and it generated 73 classification rules collectively after the elimination of the rules with zero weights (no contribution to the classification system). Few rules which are classified as no class values are characterized as null. (Table 17) shows the statistics of the final results of the proposed framework.

Table 17: Statistics on Test Set of Car Evaluation Data Set

Criteria	Results (Before Processing)	Results (After Processing)
Correctly Classified Instances	1495	1706
Incorrectly Classified Instances	64 (3.7415%)	0
Unclassified Instances	169 (9.8%)	22 (1.3%)
Total Number of Instances	1728	1728
Overall Performance (Accuracy)	86.3%	98.5%
Total Number of Generated Rules	117	73
Pruned Rules	0	44

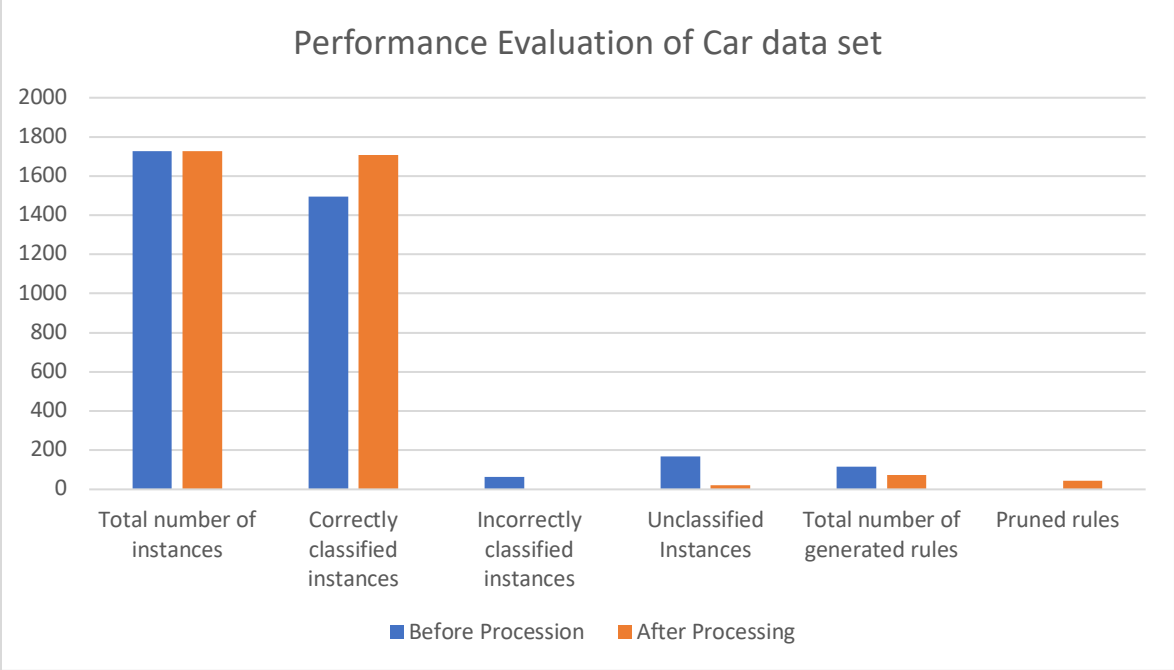


Figure 16: Performance Evaluation of Car Data Set

(Figure 16) shows the performance evaluation of car data set.

During the classification process the bucket size and range is decided. Classified rules are grouped in these buckets according the threshold value. For car evaluation data set, total 7

buckets are produced along with the range difference of 1. The privacy results generated after the processing of proposed work are shown in (Table 18):

Table 18: Privacy statistics of Car data set

Privacy Metrics	Results
Total no. of buckets	7
Total no. of sensitive rules	53
Privacy Preserved percentage	88.4
Maximum Similarity Measure	7
Minimum Similarity Measure	0
Bucket Range Difference	1
k-Anonymity	K=2

4.2.3 Balance Scale Data Set

This data set was produced to represent the results of psychological experiments. Each instance is classified as pointing the balance scale pointer to the right, the left, or to the mid-point where the balance scale is balanced. The attributes of this data sets are: "Left-Weight"', "Left-Distance", "Right-Weight", & "Right-Distance". The right method to find the target class is: left side of balance scale (left-distance * left-weight) is higher than right side or right side (right-distance * right-weight) is higher than the left side. It is balanced if both left and right weights are equal. (Table 19) shows the attribute information of the balance scale data set.

Table 19: Balance Scale Data set Attribute Information

Attribute Name	Attribute Value
Target class	L, B, R
Left Weigh	1, 2, 3, 4, 5
Left Distance	1, 2, 3, 4, 5
Right Weight	1, 2, 3, 4, 5
Right Distance	1, 2, 3, 4, 5

Few attributes are tagged as privacy parameters. (Table 20) shows the assumption made to check the validity and correctness of the designed framework.

Table 20: Balance Scale Dataset Privacy Parameters

Privacy Parameters	Tagged Attributes
Quasi Identifiers	Right Weight, Left Weight
Sensitive Attributes	Target Class

After the percentage split, 66% of the training data set is passed to multiple learning classifiers and it generated 63 classification rules collectively after the elimination of the rules with zero weights (no contribution to the classification system). Few rules which are classified as no class values are characterized as null.

(Table 21) shows the statistics of the final results of the proposed framework.

Table 21: Statistics on Test Set of Balance Scale Data Set

Criteria	Results (Before Processing)	Results (After Processing)
Correctly Classified Instances	329	457
Incorrectly Classified Instances	103 (16.6%)	168
Unclassified Instances	193 (30.9%)	0
Total Number of Instances	625	625
Overall Performance (Accuracy)	52 %	73%
Total Number of Generated Rules	111	63
Pruned Rules	0	48

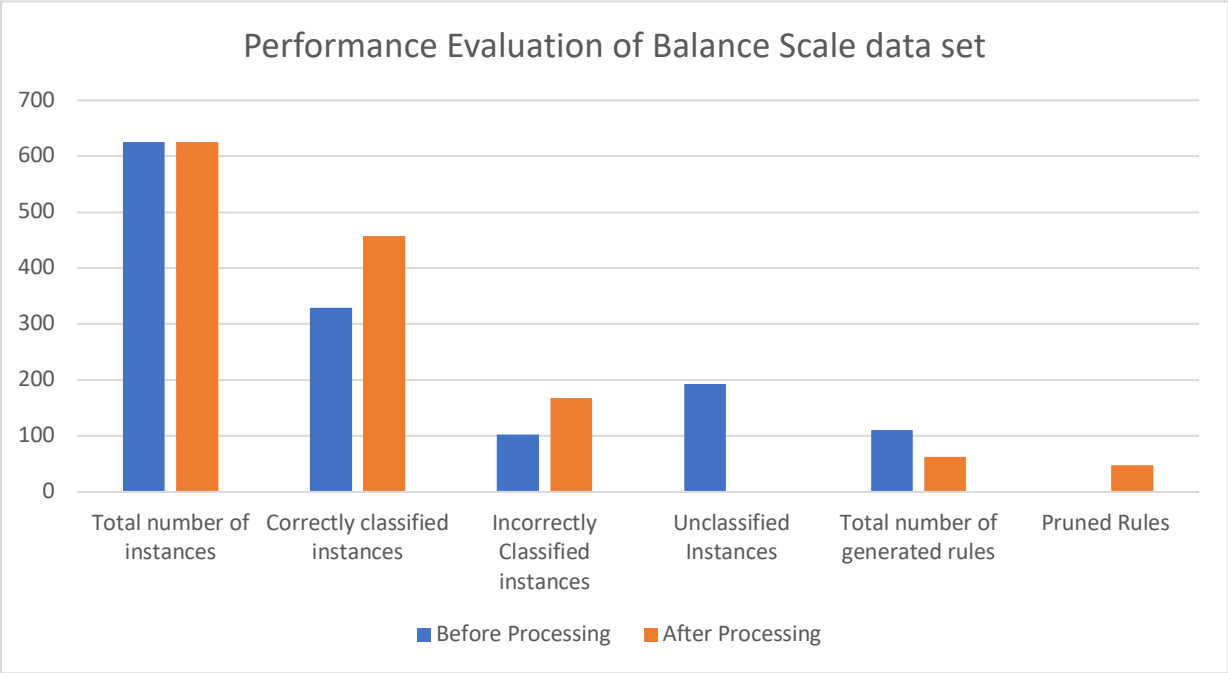


Figure 17: Performance evaluation of Balance Scale Data Set

(Figure 17) shows the performance evaluation of the balance scale data set.

During the classification process, the bucket size and range is decided. Classified rules are grouped in these buckets according to the threshold value. For the balance scale data set, a total of 4 buckets are produced along with the range difference of 1. The privacy results generated after the processing of the proposed system are following:

Table 22: Privacy statistics of Balance Scale data set

Privacy Metrics	Results
Total no. of buckets	4
Total no. of sensitive rules	32
Privacy Preserved percentage	85.2
Maximum Similarity Measure	4
Minimum Similarity Measure	0
Bucket Range Difference	1
k-Anonymity	K=2

4.2.4 Tic-Tac-Toe Data Set

This data set displays the complete feasible set of board configurations at the end of "tic-tac-toe" games, the rules of this game are: 'x' player supposed to have the first turn to play. The target conception is "win for x" (i.e., true when 'x' has 1 of 8 possible ways to create a three-in-a-row). (Table 23) shows the detailed attribute description of the Tic-Tac-Toe data set:

Table 23: Tic-Tac-Toe Data set Attributes Information

Attributes	Values
Target class	Negative, Positive
Top left square	X, O, B
Top middle square	X, O, B
Top right square	X, O, B
Middle left square	X, O, B
Middle middle square	X, O, B
Middle right square	X, O, B
Bottom left square	X, O, B
Bottom middle square	X, O, B
Bottom right square	X, O, B

"x" player means that "x" is taken &"o" player means that "o" is taken, "b" means cell or square is blank and not yet marked by any player.

Few attributes are tagged as privacy parameters. (Table 24) shows the assumption made to check the validity and correctness of the designed framework.

Table 24: Tc-Tac-Toe Dataset Privacy Parameters

Privacy Para Meters	Tagged Attributes
Quasi Identifiers	Top Right Square, Bottom Right Square
Sensitive Attributes	Target Class

After the percentage split, 66% of the training data set is passed to multiple learning classifiers and it generated 117 classification rules collectively after the elimination of the rules with zero weights (no contribution to the classification system). Few rules which are classified as no class values are characterized as null.

(Table 25) shows the statistics of the final results of the proposed framework.

Table 25: Statistics on Test Set of Tic- Tac-Toe Dataset

Criteria	Results (Before Processing)	Results (After Processing)
Total Number of Instances	958	958
Correctly Classified Instances	864	859
Incorrectly Classified Instances	59 (6.2%)	0
Unclassified Instances	53 (5.5%)	99 (10.3%)
Overall Performance (Accuracy)	88.2%	89%
Total Number of Generated Rules	200	117
Pruned Rules	0	83

Figure 18: Performance evaluation of Tic-Tac-Toe Data Set

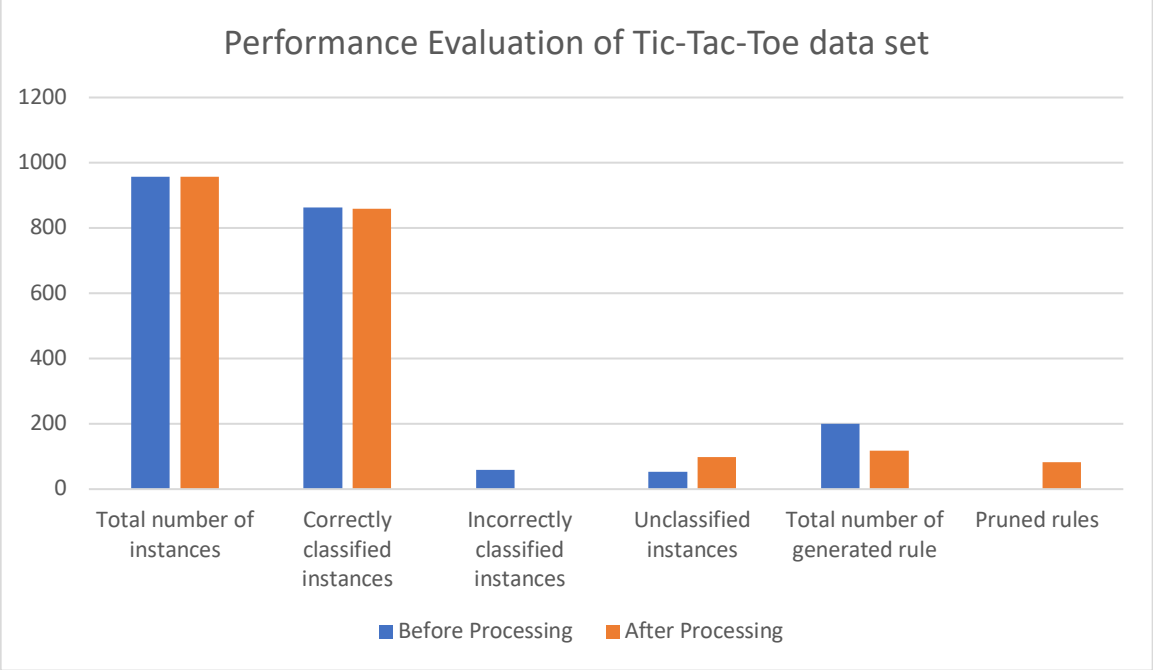


Figure 18 shows the performance evaluation of the Tic-Tac-Toe data set.

During the classification process, the bucket size and range is decided. Classified rules are grouped in these buckets according to the threshold value. For the tic-tac-toe data set, a total of 4 buckets are produced along with the range difference of 1. The privacy results generated after the processing of the proposed system are shown in (Table 26):

Table 26: Privacy statistics of Tic-Tac-Toe data set

Privacy Metrics	Results
Total no. of buckets	4
Total no. of sensitive rules	102
Privacy Preserved percentage	90.1
Maximum Similarity Measure	3
Minimum Similarity Measure	0
Bucket Range Difference	1
k-Anonymity	K=3

4.2.5 Heart Disease Data set

This data set contains information about heart patients. Total 270 observations are recorded from multiple patients. The purpose of this data set is to predict heart diseases as it defines heart diseases' risk factors. (Table 27) shows the detailed description of the heart disease data set.

Table 27: Heart Disease Dataset Attribute information

Attributes	Values
Age	29 - 27
Sex	Male , Female
Chest Pain type	Levels: 1 - 4
Resting blood pressure	94 - 200
Serum cholesterol	126 – 256 mg/dl
Fasting blood sugar	0-1(if > 120 mg then 1, otherwise 0)
Resting Electrocardiographic results	0 - 2

Maximum Heart rate achieved	71 - 202
Exercise induced angina	0 - 1
Old peak	0 - 62
The slope of the peak exercise ST segment	1- 3
Number of major vessels	0 – 3
Thal	3 – 7
Target class	Present, Absent

Few attributes are tagged as privacy parameters. (Table 28) shows the assumption made to check the validity and correctness of the designed framework.

Table 28: Privacy parameters of Heart disease dataset

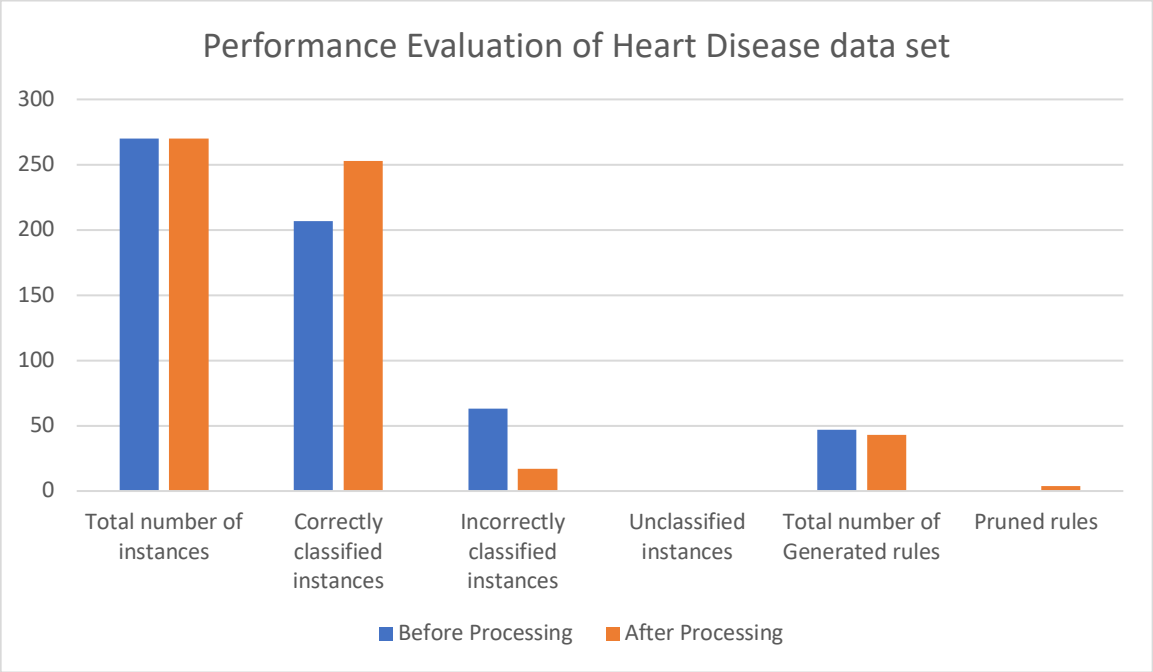
Privacy Parameters	Tagged Attributes
Quasi Identifiers	Age, Sex
Sensitive Attributes	Target Class

After the percentage split, 66% of the training data set is passed to the multiple learning classifiers and it generated 43 classification rules collectively after the elimination of the rules with zero weights (no contribution to the classification system). Few rules which are classified as no class values are characterized as null. Table 29 shows the statistics of the final results of the proposed framework:

Table 29: Statistics on test set of Heart Disease dataset

Criteria	Results (Before Processing)	Results (After Processing)
Total Number of Instances	270	270
Correctly Classified Instances	207	253
Incorrectly Classified Instances	63 (23.33%)	17(6.30 %)
Unclassified Instances	0	0
Overall Performance (Accuracy)	76.66%	89%
Total Number of Generated Rules	47	43
Pruned Rules	0	4

Figure 19: Performance evaluation of Heart Disease Data Set



(Figure 19) shows the performance evaluation of heart disease data set.

During the classification process, the bucket size and range is decided. Classified rules are grouped in these buckets according to the threshold value. For the heart disease data set, a total of 4 buckets are produced along with the range difference of 1. The privacy results generated after the processing of the proposed system are shown in (Table 30):

Table 30: Privacy statistics of Heart Disease dataset

Privacy Metrics	Results
Total no. of buckets	4
Total no. of sensitive rules	20
Privacy Preserved percentage	73.5
Maximum Similarity Measure	4
Minimum Similarity Measure	0
Bucket Range Difference	1
k-Anonymity	K=2

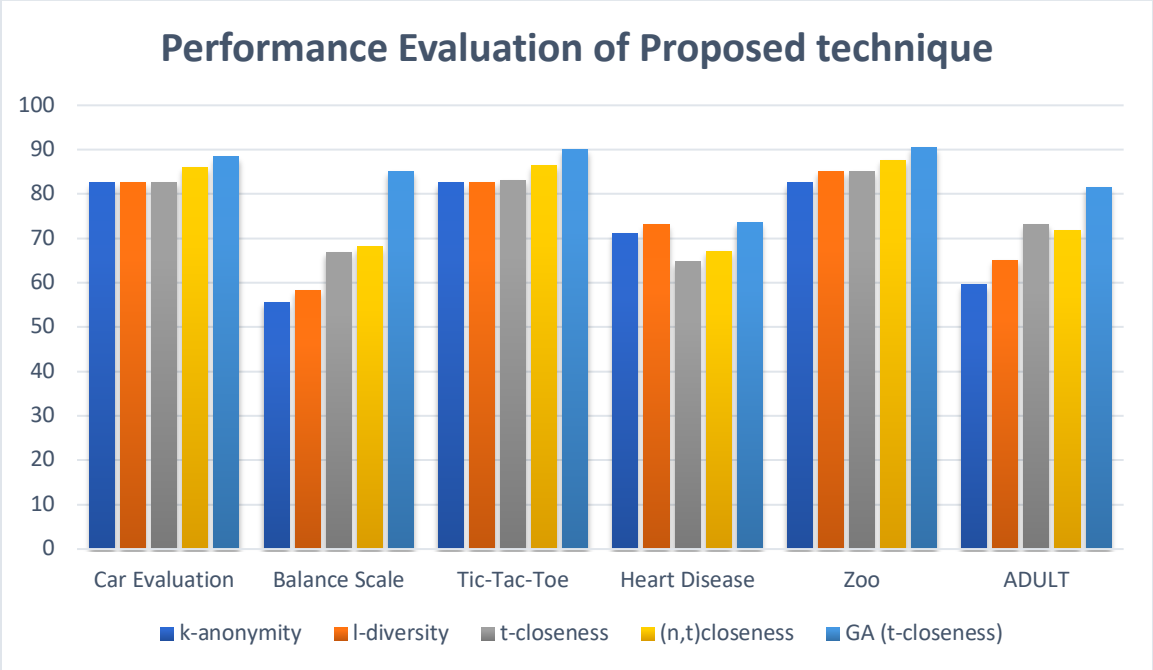
4.3 Comparative Analysis of Proposed Technique with other Privacy Approaches

Different privacy-preserving techniques generate different results based on given test data sets. Privacy of those test data sets are measured using the presence of the sensitive rules and the privacy technique used to hide sensitive information. The following results showing the comparative analysis of the developed technique with all other privacy techniques used to see the performance of the proposed work in compliance with them.

Table 31: Comparative Analysis of developed technique with other privacy techniques w.r.t. privacy preservation

Data Sets	Privacy Techniques				
	k-Anonymity	l-Diversity	t-closeness	(n,t) Closeness	GA(t-closeness)
Car Evaluation	82.4	82.6	82.6	86.1	88.4
Balance Scale	55.5	58.2	66.9	68.1	85.2
Tic-Tac-Toe	82.5	82.6	83.1	86.5	90.1
Heart Disease	71.1	73.2	64.8	67.1	73.5
Zoo	82.5	85.1	85.2	87.6	90.5
ADULT	59.6	65.1	73.2	71.9	81.5

Figure 20: Comparative Analysis of Developed Technique with other Privacy Approaches



(Table 32) shows comparative analysis of classification accuracy of developed technique with the other techniques.

Table 32: Comparative analysis of classification accuracy of developed technique

	Single classifier (Decision tree)	Multiple Classifiers
ADULT	86%	98.6%
Heart Disease	76.66	89%
Car Evaluation	86.3%	98.5%
Balance Scale	52%	73%
Tic-Tac-Toe	88.2%	89%
Zoo	92.7	94.17

Chapter 5

Conclusion and Future Work

5 Conclusion

In this varied field of privacy-preserving data mining (PPDM), many researchers are working on making data secure using different data protection techniques like generalization and data distortions, and then forward it to the mining process for data analysis by applying data mining techniques on anonymized data. Some researchers are working opposite of it; first, they apply mining techniques then they apply different privacy-preserving techniques to anonymize the data. A trade-off is required between the accuracy and privacy in both methods.

- My research work belongs to the research areas where classification and privacy-preserving techniques work together to generate the final results in the form of anonymized rules and instances while maintaining the correlations of key and sensitive attributes.
- Taking into consideration that data distortion needs to be controlled according to the required amount of applicability.
- A framework is proposed to address the concerns related to data privacy.
- This framework has enough flexibility that it can be made compliant with all types of datasets, datasets must be according to WEKA defined formats [58] [56].
- It is also considered that participating parties agree on defining the level of sensitivity of attributes and taxonomy for categorical data types. It needs to pass through different steps to make it useful as well as private.
- The threshold value of t-closeness is optimized using GA to maximize the data utility and privacy.

Classification accuracy is increased with the combination of rules generated from three the best learning classifiers. This framework is very much beneficial in the field of medicine, where there are multiple data owners and sites are working.

- Data is gathered from different hospitals and analyzed for further follow-ups of the patients and medical researches.

Hospitals need to give special attention to preserving privacy by over-communicating sensitive information. This research is a step towards building a better society by focusing on the essential concern of privacy over large information, especially in a distributed environment.

5.1 Future Works

In the later versions of this framework, the following enhancements are possible for the researchers in this domain to work on:

- The proposed approach is designed based on distributions in the equivalence classes, semantics of data should be considered as well for finding inter/intra cross similarities among the attributes for better results.
- The threshold value of t-closeness can be optimized through other evolutionary algorithms like the artificial neural network to get better results.
- Currently, the privacy parameters are provided as input. The component can be added that entertains the international HIPPA (Health Insurance Portability and Accountability Act) rules about the sensitivity of the information to be hidden or suppressed.

References

References

[1] Fraser, A., Monte Carlo analyses of genetic models. Nature, 1958. 181(4603): p. 208-209.

[2] Holland, J.H., Adaptation in natural and artificial systems. An introductory analysis with application to biology, control, and artificial intelligence. Ann Arbor, MI: University of Michigan Press, 1975.

[3] Cohen, W.W. Fast effective rule induction. in Proceedings of the twelfth international conference on machine learning. 1995.

[4] Frank, E. and I.H. Witten, Generating accurate rule sets without global optimization. 1998.

[5] Quinlan, J.R. and R.M. Cameron-Jones. C 4. 5: Programs for Machine Learning. in European conference on machine learning. 1993. Springer.

[6] Funatsu, K. and K. Hasegawa, New fundamental technologies in data mining. First published January, 2011.

[7] Aggarwal, C.C. and S.Y. Philip, Privacy-Preserving Data Mining: A Survey, 2008, Springer.

[8] Warner, S.L., Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 1965. 60(309): p. 63-69.

[9] Dasseni, E., et al. Hiding association rules by using confidence and support. in International Workshop on Information Hiding. 2001. Springer.

[10] Liu, K., C. Giannella, and H. Kargupta, An attacker's view of distance preserving maps for privacy preserving data mining. Knowledge Discovery in Databases: PKDD 2006, 2006: p. 297-308.

[11] Donoho, D.L., High-dimensional data analysis: The curses and blessings of dimensionality. AMS Math Challenges Lecture, 2000. 1: p. 32.

[12] Evfimievski, A., J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. in Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. 2003. ACM.

[13] Li, F., et al. Hiding in the crowd: Privacy preservation on evolving streams through correlation tracking. in Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. 2007. IEEE.

[14] Fienberg, S.E. and J. McIntyre. Data swapping: Variations on a theme by dalenius and reiss. in Privacy in statistical databases. 2004. Springer.

[15] Sweeney, L., k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002. 10(05): p. 557-570.

[16] LeFevre, K., D.J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. in Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on. 2006. IEEE.

[17] Machanavajjhala, A., et al. l-diversity: Privacy beyond k-anonymity. in Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on. 2006. IEEE.

[18] Li, N., T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. in Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. 2007. IEEE.

[19] Naboulsi, D., et al., Large-scale mobile traffic analysis: a survey. IEEE Communications Surveys & Tutorials, 2016. 18(1): p. 124-161.

[20] Sei, Y., et al., Anonymization of Sensitive Quasi-Identifiers for l-diversity and t-closeness. IEEE Transactions on Dependable and Secure Computing, 2017.

[21] Gkoulalas-Divanis, A., G. Loukides, and J. Sun, Publishing data from electronic health records while preserving privacy: A survey of algorithms. Journal of biomedical informatics, 2014. 50: p. 4-19.

[22] Yang, J.-J., J.-Q. Li, and Y. Niu, A hybrid solution for privacy preserving medical data sharing in the cloud environment. Future Generation Computer Systems, 2015. 43: p. 74-86.

[23] Nagaraju, S. and L. Parthiban, Trusted framework for online banking in public cloud using multi-factor authentication and privacy protection gateway. Journal of Cloud Computing, 2015. 4(1): p. 22.

[24] Sim, T. and L. Zhang. Controllable face privacy. in Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on. 2015. IEEE.

[25] Nayak, G. and S. Devi, A survey on privacy preserving data mining: approaches and techniques. International Journal of Engineering Science and Technology, 2011. 3(3).

[26] Singh, U.K., B.K. Pandya, and K. Dixit, An Overview on Privacy Preserving Data Mining Methodologies. International Journal of Engineering Trends and Technology, 2011.

[27] Adam, N.R. and J.C. Worthmann, Security-control methods for statistical databases: a comparative study. ACM Computing Surveys (CSUR), 1989. 21(4): p. 515-556.

[28] Atallah, M., et al. Disclosure limitation of sensitive rules. in Knowledge and Data Engineering Exchange, 1999.(KDEX'99) Proceedings. 1999 Workshop on. 1999. IEEE.

[29] Oliveira, S., O. Zaïane, and Y. Saygin, Secure association rule sharing. Advances in Knowledge Discovery and Data Mining, 2004: p. 74-85.

[30] Aggarwal, C.C., J. Pei, and B. Zhang. On privacy preservation against adversarial data mining. in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006. ACM.

[31] Chang, L. and I.S. Moskowitz. Parsimonious downgrading and decision trees applied to the inference problem. in Proceedings of the 1998 workshop on New security paradigms. 1998. ACM.

[32] Moskowitz, L. and I.S. Chang, A decision theoretical based system for information downgrading, 2000, NAVAL RESEARCH LAB WASHINGTON DC CENTER FOR HIGH ASSURANCE COMPUTING SYSTEMS (CHACS).

[33] Dobkin, D., A.K. Jones, and R.J. Lipton, Secure databases: Protection against user influence. ACM Transactions on Database systems (TODS), 1979. 4(1): p. 97-106.

[34] Kenthapadi, K., N. Mishra, and K. Nissim. Simulatable auditing. in Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. 2005. ACM.

[35] Dwork, C., et al. Calibrating noise to sensitivity in private data analysis. in TCC. 2006. Springer.

[36] Nabar, S.U., et al. Towards robustness in query auditing. in Proceedings of the 32nd international conference on Very large data bases. 2006. VLDB Endowment.

[37] Natwichai, J., X. Li, and M.E. Orlowska. A reconstruction-based algorithm for classification rules hiding. in Proceedings of the 17th Australasian Database Conference-Volume 49. 2006. Australian Computer Society, Inc.

[38] Kam, J.B. and J.D. Ullman, A model of statistical database their security. ACM Transactions on Database Systems (TODS), 1977. 2(1): p. 1-10.

[39] Kleinberg, J., C. Papadimitriou, and P. Raghavan, Auditing boolean attributes. Journal of Computer and System Sciences, 2003. 66(1): p. 244-253.

[40] Dwork, C., et al. Our Data, Ourselves: Privacy Via Distributed Noise Generation. in Eurocrypt. 2006. Springer.

[41] Agrawal, R., et al. Auditing compliance with a hippocratic database. in Proceedings of the Thirtieth international conference on Very large data bases-Volume 30. 2004. VLDB Endowment.

[42] Chawla, S., et al. Toward Privacy in Public Databases. in TCC. 2005. Springer.

[43] Reiss, S.P., Security in databases: A combinatorial study. Journal of the ACM (JACM), 1979. 26(1): p. 45-57.

[44] Dwork, C. and M. Naor, On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. Journal of Privacy and Confidentiality, 2008. 2(1): p. 8.

[45] Mishra, N. and M. Sandler. Privacy via pseudorandom sketches. in Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. 2006. ACM.

[46] Denning, D.E., Secure statistical databases with random sample queries. ACM Transactions on Database Systems (TODS), 1980. 5(3): p. 291-315.

[47] Blum, A., et al. Practical privacy: the SuLQ framework. in Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. 2005. ACM.

[48] Hore, B., S. Mehrotra, and G. Tsudik. A privacy-preserving index for range queries. in Proceedings of the Thirtieth international conference on Very large data bases-Volume 30. 2004. VLDB Endowment.

[49] Clifton, C., et al., Tools for privacy preserving distributed data mining. ACM Sigkdd Explorations Newsletter, 2002. 4(2): p. 28-34.

[50] Valentini, G. and F. Masulli. Ensembles of learning machines. in Italian Workshop on Neural Nets. 2002. Springer.

[51] Dietterich, T.G., Ensemble methods in machine learning. Multiple classifier systems, 2000. 1857: p. 1-15.

[52] Duin, R.P. and D.M. Tax. Experiments with classifier combining rules. in

International Workshop on Multiple Classifier Systems. 2000. Springer.

[53] Kittler, J., et al., On combining classifiers. IEEE transactions on pattern analysis and machine intelligence, 1998. 20(3): p. 226-239.

[54] Breiman, L., Bagging predictors. Machine learning, 1996. 24(2): p. 123-140.

[55] Au, W.-H., K.C. Chan, and X. Yao, A novel evolutionary data mining algorithm with applications to churn prediction. IEEE transactions on evolutionary computation, 2003. 7(6): p. 532-545.

[56] Blake, C.L. and C.J. Merz, UCI Repository of machine learning databases [http://www. ics. uci. edu/~ mlearn/MLRepository. html]. Irvine, CA: University of California. Department of Information and Computer Science, 1998. 55.

[57] Rubner, Y. and C. Tomasi, The earth mover's distance, in Perceptual Metrics for Image Database Navigation. 2001, Springer. p. 13-28.

[58] Garner, S.R. Weka: The waikato environment for knowledge analysis. in Proceedings of the New Zealand computer science research students conference. 1995.

[59] J. Bauto, R. Neves and N. Horta, Parallel Genetic Algorithms for Financial

Pattern Discovery using GPUs. 2018 , Springer.

[60] Frances Bountempo edited by Tammy Coron, Genetic Algorithms and Machine Learning for Programmers (Create AI Models and Evolve Solutions) [1 ed.], 2019.

[61] Alexandre Bergel, Agile Artificial Intelligence in Pharo: Implementing Neural Networks, Genetic Algorithms and Neuroevolution, 2020.

[62] P. Christen, T. Ranbaduge, R. Shnell, Linking Sensitive Data: Methods and Techniques for Practical Privacy-Preserving Information Sharing, Springer 2020.