The Seybold Report

ISSN 1533-9211

DOI 10.5281/zenodo.7374949

NOVEL RAENOPTI APPROACH FOR SECURE AND PRIVACY PRESERVING CLASSIFICATION OF HORIZONTALLY PARTITIONED MENTAL HEALTH DATA IN DISTRIBUTED ENVIRONMENT

VIJAYA PINJARKAR[*1], AMIT JAIN[2], ANAND BHASKAR[3] and ASHUTOSH GUPTA[4]

[1]Department of Information Technology, K.J. Somaiya Institute of Engineering and Information Technology, Mumbai, Maharashtra, India. *Corresponding Author Email: *[1]vkhirodkar@somaiya.edu

[1, 2, 3, 4] Department of Computer science and Engineering, Sir Padampat Singhania University, Udaipur, Rajasthan, India. Email: [2]amit.jain@spsu.ac.in, [3]anand.bhaskar@spsu.ac.in, [4]ashu.gupta@spsu.ac.in

Abstract:

To carry out any research the first step is data gathering, as accuracy and prediction is depending on data. In the digital era, data generated at different sectors like banking, education, healthcare center or hospitals etc. is numerous which is helpful for researcher to carry out research. However, sharing data in plain text format, may compromise privacy of data. So, to maintain the privacy and secrecy of data this paper proposes RaEnOpti Approach, where optimized and secured data will be shared with researcher. This approach uses privacy preserving data mining, Encryption techniques and Genetic Algorithm for optimization work. This paper is discussing about the work in distributed network, as the data is collected from different healthcare centers.

Keywords: Privacy, Secrecy, Genetic Algorithm, Distributed Network.

1) INTRODUCTION

The whole world is connected with internet and using different technologies. All smart devices are full with different gadgets of different Apps like, games, shopping apps, education apps, and healthcare apps etc. uses of devices generates huge data. As well uses of technology in different sectors like education, healthcare, medical, financial sector generates huge data. Meaningful data is every time useful for various kind of research in germane field. For extraction of meaningful data from such databases, data mining techniques are always used. At the same time extracted information is helpful to doctors, medical staff, therapists for decision making like choosing the correct treatment, picturing diseases on new cases in fewer time [1]. On the other hand, researchers are starving to work on real time data. There are websites where datasets are present but that are limited one.

The need for mental health is as strong as the importance of physical health in the world after the Corona global catastrophe right now we are living in a different age. We think that we should be present everywhere. We should be connected with everyone. But social media has made us anti-social which is doing more harm than our expectations. Staying is avoidance. As of the several areas of health care, the last two decades have found several cases related to mental health [2]. Mental health does not constantly stay similar all the way through human life. It can change as circumstances deviations and as one move through different phases of life. There is a humiliation attached to the mental health problems. This means that people feel

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7374949

uncomfortable about them and do not talk about it much. Many time people do not feel comfortable to talk about their feelings. But it is healthy to know and say how you are feeling. Hence, the work focuses to contribute towards mental health care.

The current era has seen data mining and machine learning techniques producing excellent outcomes for diverse applications using the data generated by different applications. Data related to mental health is available with doctors or counsellors but it is highly confidential data so doctors or counsellors are not ready to share this type of data as it concerns about the privacy of the underlying patients. So, to address the issue of privacy the upcoming variant of data mining like privacy preserving data mining (PPDM)[3] can be used.

When the health care center give data to researcher, this action is called as data transfer. When the data transfer from one party to another party, there may be chances of confidentiality compromise on the channel. So cryptographic approach can be used to achieve confidentiality. Cryptography is creating written or generated codes that allows information to be kept undisclosed[4]. It is the exercise and study of techniques for securing communication and data in the presence of challengers.

Once all the data has been successfully decrypted and collected at centralized server it is very much important to segregate the data and remove the duplicate data, for smooth execution of this task Genetic Algorithm (GA) is used. There are different techniques available for optimization but GA are robust and provide optimization over large space state and unlike traditional AI, GA do not break on small changes in input or presence of noise in input.

This paper mainly discusses about application of Privacy Preserving Data Mining and Cryptography on mental health data which is collected from different healthcare centers and optimized using Genetic Algorithm so that researchers should get real data to work and doctors get optimized and meaningful data which ultimately result in less diagnose time and less expenditures on different reports and medicines.

Paper is divided in to Section-I Introduction, Section-II, and Literature survey, Section-III, Existing System, Section-IV, proposed system, Section-V, Implementation details, Section-VI, Observations and Finally conclusion and references included in paper.

2) LITERATURE SURVEY

As the research work comprises three areas literature survey also consist of them as A. Privacy Preserving Data Mining, B. Cryptography, and C. Optimization using Genetic Algorithm.

A. Privacy Preserving Data Mining (PPDM):

PPDM is a submission of data mining result in response to privacy security in data mining. PPDM is also called a privacy enhancing or privacy sensitive data mining. It deals with obtaining true data mining without disclosing the basic sensitive data values. Privacy preservation in data mining has emerged as a prerequisite for trading confidential information in term of data analysis, validation publication, etc. The field of privacy has seen swift advance in recent years as there is growth in ability to store data. The author in [3] [5] introduced the

ISSN 1533-9211

DOI 10.5281/zenodo.7374949

privacy preserving data mining. The PPDM is mainly classified in two types:

- Based on the data lifecycle phase: the data lifecycle phase at which the privacy preservation ensured in data collection, data publishing, data distribution and the output of the data mining. Table 1 shows the data publishing privacy techniques taxonomy.

Table 1: Data Publishing Privacy Techniques Taxonomy

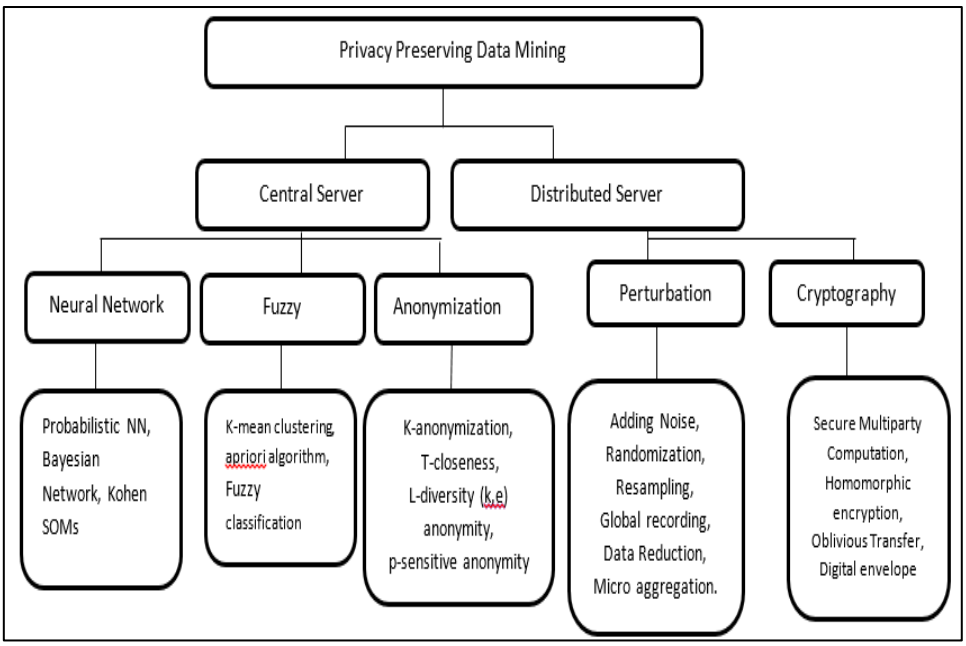
Privacy Model	Methods	Description
k-anonymity [6]	Generalization, suppression	Anonymity is guaranteed by the existence of at least other k-l undistinguishable records for each in a database. This group of k Undistinguishable records are referred to as equivalence class.
l-diversity [7]	Generalization, Suppression	Expands the k-anonymity model by requiring every equivalence class to have at least I "well represented" values for the sensitive attributes.
t-closeness [8]	Generalization, Suppression	Solve the l-diversity problem of twisted sensitive values distribution by requiring that the distribution of the sensitive values in each equivalence class to be "close" to the corresponding distribution in the original table, where close means upper bounded by a threshold t.
Personalized privacy [9]	Generalization	Achieved by creating a classification tree using generalization, and by allowing the record owners to define a guarding node. Owners' privacy is penetrated if an attacker is allowed to infer any sensitive value from the sub-tree of the guarding node with a probability (breach probability) greater than a certain threshold.

- Based on the location of computation: Based on the location of computation carried out for mining results, PPDM techniques is classified as described in figure1. Classification is broadly categorized as: central and distributed server.

ISSN 1533-9211

DOI 10.5281/zenodo.7374949

Figure1: Classification of Privacy Preserving Data Mining algorithm.



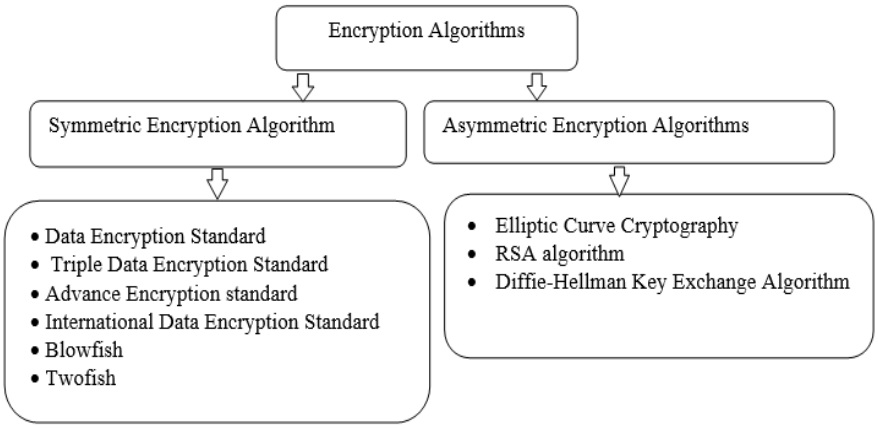
Authors in [10] shows the comparison of perturbation methods.

B. Cryptography:

Cryptography is the process which allows secure transition of information, by maintaining confidentiality and integrity of information [11] while confidentiality is defined as protecting information from being accessed by unauthorized user.

Figure 2, shows the classification of cryptographic approach. Based on the key used they are categorized as a) Symmetric encryption and b) asymmetric encryption.

Figure2: Classification of Encryption Algorithms



ISSN 1533-9211

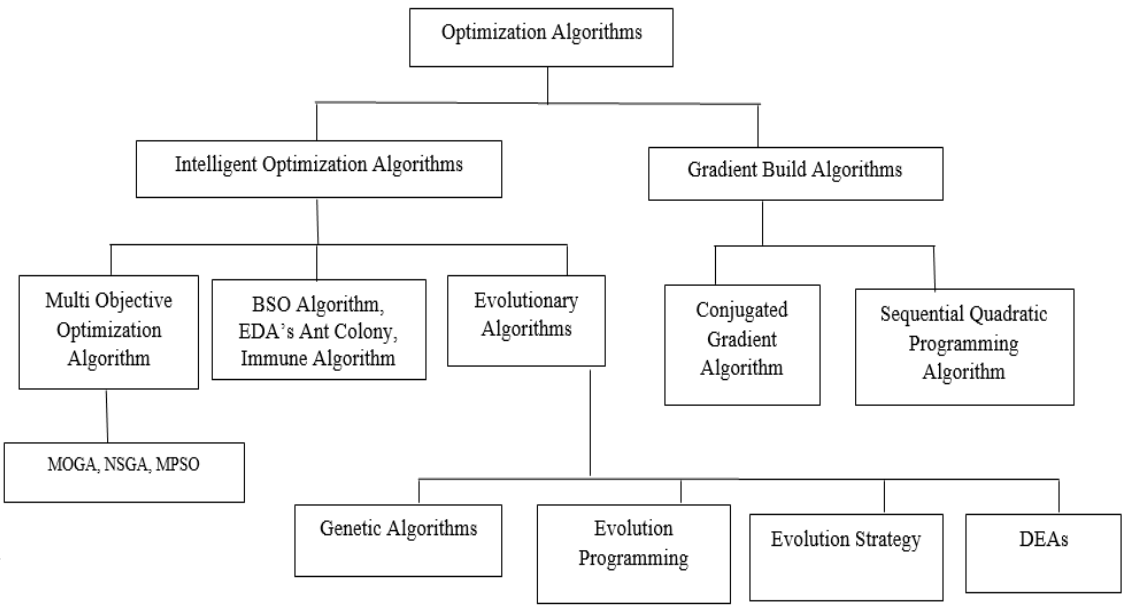
DOI 10.5281/zenodo.7374949

Cryptograph is used in many applications for providing secrecy like hashing algorithms are used in many applications for storing password. In [12] author discussed the homomorphic cryptography and pairing-based cryptography. [13] Compared and analyzed various homomorphic encryption algorithms. In some application GPS in location privacy encryption algorithms are very useful [14]. Paper [15] shows the comparison different symmetric encryption algorithms.

C. Optimization using Genetic Algorithm (GA):

GA is utmost commonly used in optimization problem wherein objective functions value is maximize or minimize under a given set of constrains. In Genetic Algorithm's there is pool of possible solution to the given problem. These solutions then undergo recombination and mutation, producing new children, the process is repeated over various generations. The genetic algorithm is a technique for solving both constrained and unconstrained optimization problems that is based on natural selection, the process that drives biological evolution. GA is a search optimization technique based on the principles of genetics and natural selection. Various optimization techniques like Genetic Algorithm, Ant Colony Optimization[16],Particle Swarm Optimization[17] etc. are available which can be applied for getting efficient results. The [18] and [19] presented the comparative study of the optimization algorithm. Optimization techniques are applied in several things, like [20] presented a survey for optimization techniques in data management. Figure3, shows the classification of genetic algorithms. Optimization algorithm principally divided into two parts Intelligent Optimization Algorithm (IOA) and Gradient Build Algorithms (GBA). Work in this paper focused on GA which is evolutionary from intelligent optimization algorithm.

Figure 3: Classification of optimization algorithm in machine learning



The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7374949

Table 2, shows study of different papers where genetic algorithms are used for optimization.

Table 2: GA optimization application in exiting systems

Sr. No	Authors	Reference no	Year of Publication	Application area	Remarks
1	Manish Saggaar, Ashish Kumar Agrawal and Abhimanyu Lad	[21]	2004	Optimization of Association Rule Mining	The rules generated by association rule mining do not consider the negative occurrences of attributes, but by using GA's over these rules the system can predict the rules contains negative attributes.
2	Akinori Hirabayashi, Claus Aranha and Hitoshi Iba	[22]	2009	Optimization of trading rule in foreign exchange.	By Applying a variety of Technical Indexes, GA managed to find the most Profitable trading rule for the period was searched by GA. This rule was applied in the time period immediately posterior without being re-trained beforehand. As a result, though there is vulnerability to sudden market changes, the effectiveness of the proposed method was shown in some aspects.
3	Xiang Zhang, Huaixiang Zhang, Ertao Li and Xiang Zhang	[23]	2010	Optimization of Fuzzy Classification System	A novel approach to construct FCS using CA algorithm and GA has been proposed according to the relation between the performance of FCS and accuracy and Interpretability.
4	Yan, Lewei Sun, Zuoyu Mao, Keyang	[24]	2013	Robust optimization	A robust GA which performs random perturbation during optimization processes has been applied to some mathematical problems to show that it works as fast as the usual GAs.
5	Hassoon, Mafazalyaqeen Kouhi, Mikhak Samadi Zomorodi-Moghadam, Mariam Abdar Moloud	[25]	2017	Rules Optimization for Liver disease Prediction.	Presents an optimization approach to reduce and optimize the rules of disease diagnosis.
6.	Interciso MateusGarcia, Plínio Barrio	[26]	2020	Genetic Algorithm for optimizing stock usage	This is common supply chain problem. Very efficient use of Genetic Algorithm in optimizing stock usage in supply chain problem.

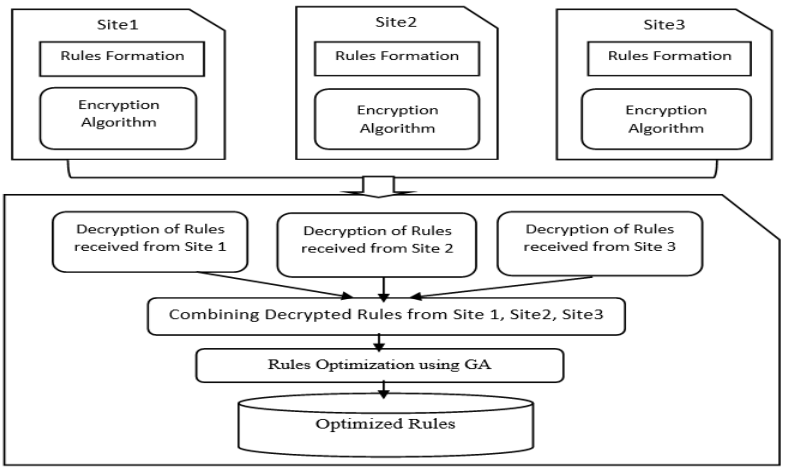
ISSN 1533-9211

DOI 10.5281/zenodo.7374949

3) PROPOSED SYSTEM

This section discusses about the Novel RaEnOpti Approach, which is shown in figure 4. The principle aim of this approach is to remove the barrier between researchers and data owners where privacy and confidentiality of data is maintained. Data owners mean sectors like banking, education, healthcare centre or hospitals etc. In this paper the discussion and implementation is around mental health datasets, where the privacy and secrecy is on topmost priority. Different hospitals / counsellors maintain mental health patient data, from figure4, site1, 2, 3 represent different hospitals / counsellors which collect patient details like name, age, gender, occupation, family details, previous history if any, work culture, type of disease etc. Once the data is collected it gets pre-process. Pre-processed data is then encoded, and PPDM technique is applied on it. Once the data is perturbed the privacy of the data achieved. Perturbed data is given as input to the rule based classifier to generate rules. Next step is to apply the encryption technique on database where all rules are stored to achieve confidentiality of data. The encrypted data from different hospitals / counsellors is now handover to central unit, where firstly it gets decrypted and then combined. The combined rule dataset may contain the duplicate as well weak rules so for finding fittest rule from datasets optimization techniques is applied. Out of different available optimization techniques as discuss in section II, Genetic Algorithm is used. Such optimised database is given to researcher to carry out research which is useful for doctors to treat patient in less time and in less expenditure on different pathological tests. Figure 4, shows the system architecture.

Figure 4: System Diagram for RaEnOpti



4) IMPLEMENTATION DETAILS

For implementation of proposed system, considered mental health dataset. Mental health patient dataset for this implementation is taken from online available dataset, provided by OSMI survey. The dataset mainly covers information like age, gender, location of work, type

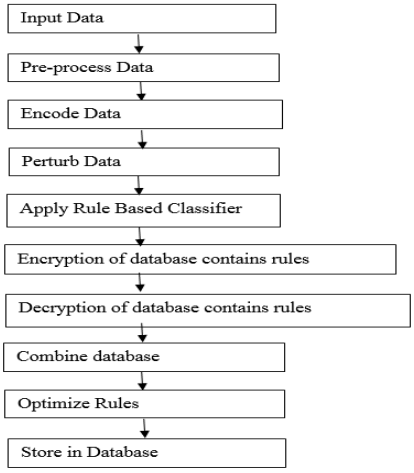
ISSN 1533-9211

DOI 10.5281/zenodo.7374949

of work, is he/she self-employed, total 27 questions are included. Sensitive attribute whose privacy is critical is considered for implementation. Implementation is done in python.

The flow of proposed system is given in below figure 5.

Figure 5: System flow diagram



Client-Side Steps:

Step 1: Pre-processing data, dataset is in .csv format. The pre-processing the first step carried out on dataset in Jupyter Notebook. Data cleaning is the practice by which data gets rid of from not required data, making it appropriate for further analysis.

Step 2: The pre-processed data is then encoded as shown in figure 6.

Step 3: Apply proposed perturbation algorithm on encoded data.

Figure 6: Result of Step 2. Screenshot of encoding

Gender	'female', 'male', 'trans'	0, 1, 2
self_employed	'Yes','No'	1, 0
family_history	'Yes','No'	1, 0
treatment	'Yes','No'	1, 0
work_interfere	"Don't know", 'Never', 'Often', 'Rarely', 'Sometimes'	0, 1, 2, 3, 4
no_employees	'01-05', '100-500', '26-100', '500-1000', '06-25', 'More than 1000'	
remote_work	'No', 'Yes'	0,1
tech_company	'No', 'Yes'	0,1
benefits	"Don't know", 'No', 'Yes'	0, 1, 2
care_options	'No', 'Not sure', 'Yes'	0, 1, 2
wellness_program	"Don't know", 'No', 'Yes'	0, 1, 2
seek_help	"Don't know", 'No', 'Yes'	0, 1, 2
anonymity	"Don't know", 'No', 'Yes'	0, 1, 2
leave	"Don't know", 'Somewhat difficult', 'Somewhat easy', 'Very difficult', 'Very easy'	0, 1, 2, 3, 4
mental_health_consequence	'Maybe', 'No', 'Yes'	0, 1, 2

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7374949

Step 4: Once the data is perturbed apply rule-based classifier on data and generate rules. Store the generated rules in one database. Sample rule set is shown in figure 7.

Figure 7: Result of step 4. Screenshot of generated rules

have a mental health disorder <= 2 AND obs_consequence <= 0 AND seek_help <= 11 AND Gender <= 4 AND phys_health_consequence <= 15: Positive-Detected (4.0)

have a mental health disorder <= 2 AND obs_consequence <= 0 AND seek_help <= 11 AND leave <= 16 AND self_employed <= 1 AND phys_health_consequence > 15 AND mental_health_interview > 2 AND Gender > 4 AND remote_work <= 7 AND mental_vs_physical > 0 AND Age <= 16: Negative-Detected (9.0)

mental_health_interview > 3 AND leave <= 16 AND tech_company > 7 AND seek_help > 11 AND remote_work > 7: Negative-Detected (7.0/1.0)

mental_health_interview > 3 AND remote_work <= 7 AND self_employed <= 1 AND obs_consequence <= 0 AND mental_health_consequence > 14 AND mental_vs_physical <= 0: Positive-Detected (6.0/1.0)

have a mental health disorder <= 2 AND obs_consequence <= 0 AND seek_help <= 11 AND phys_health_interview <= 4 AND self_employed <= 1 AND family_history > 0 AND leave <= 15 AND mental_health_interview > 2 AND Gender > 4 AND mental_health_consequence <= 14: Positive-Detected (5.0/1.0)

have a mental health disorder <= 2 AND obs_consequence <= 0 AND seek_help <= 11 AND phys_health_interview <= 4 AND self_employed <= 1 AND mental_vs_physical <= 0 AND anonymity <= 13 AND benefits <= 9 AND treatment <= 0: Negative-Detected (16.0/2.0)

Step 5: Encryption of database using symmetric key encryption algorithm at client side (doctor/counsellor)

Server-Side Steps:

Step 6: Decryption of database contains rules

Step 7: Combine all received databases and for new database

Step 8: Apply Genetic Algorithm to optimize rules, resulting in deletion duplicate as well as weak rules.

Step 9: Store optimized rules in database.

5) OBSERVATIONS

Previous section discussed about the implementation of proposed method; this section lets discuss the results. As the system has three main parts so the results and observations are also discussed part wise. Part A will discuss about the result of PPDM method. Part B will discuss about the results of encryption techniques and Part C will take care of results of optimization techniques.

Part A: PPDM is a use of data mining research in response to privacy security in data mining. As discussed in part-II there are different methods to achieve privacy and security in data

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7374949

mining. Out of different methods research work mainly focused on perturbation method. In perturbation method arbitrary noise from a known distribution is mixed to the sensitive data before the data is sent to the data miner. There two famous methods of perturbation viz additive perturbation and multiplicative perturbation. This part will discuss the proposed perturbation method.

Table3, shows the comparison of proposed perturbation method with Additive perturbation and multiplicative perturbation method.

Table 3: Comparison of Proposed perturbation method with Additive perturbation and multiplicative perturbation method

Database	Mean	Standard Deviation
Original Dataset	14.042	7.157
Additive perturbation method	18.042	7.1536
Multiplicative perturbation method	28.084	14.313
Proposed perturbation method	14.00	7.150

To denote minimal information loss, the mean of the perturbed dataset should be nearby to the mean of the original dataset. From table3, mean of Additive perturbation method, Multiplicative perturbation method is larger than the mean of Original dataset which mean there is data loss when we implement the Additive perturbation method, Multiplicative perturbation method. Proposed perturbation method mean is nearer to mean of original dataset, which means there is less data loss in the proposed perturbation method.

Standard Deviation is a statistical term used to extent the amount of dispersion around an average. Technically it is amount of volatility. Dispersion is the difference between the actual and the average value. The larger this dispersion or variability is, the higher is the standard deviation. Standard deviation of Additive perturbation and proposed perturbation method is referring closely to the standard deviation of original dataset. The proposed perturbation method is weightier than the existing methods. So proposed perturbation method is applied on pre-processed database. After perturbation the rule-based classifier is applied on perturbed data and rules are deposited in database.

Part B: Encryption is the process by which information is covered into secret code that hides the information's true meaning. The science of encryption and decryption of information is called cryptography. Encryption is a well-known technique for preserving the confidentiality of sensitive information. [27] All the generated rules are stored in database and database is encrypted with symmetric encryption algorithm. Following table shows the comparison of different symmetric encryption algorithm and also discussed the use of best selected in the proposed system. Table4, shows the comparison of DES, 3DES, AES-128, AES-192, AES-165, IDEA based on encryption time, decryption time and memory required for different file size.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7374949

Table 4: Comparison of DES, 3DES, AES-128, AES-192, AES-165, IDEA based of encryption time, decryption time and memory required

Sr No.	Algorithm Name	Plain text files in size. (kb)	Key size (bits)	Encryption time (milli-seconds)	Encrypted file size in kb	Decryption time (milli-seconds)	Decrypted file size in kb	Memory required for encryption	CPU usage in %	Cryptanalysis
1	DES algorithm	25kb	64	0.89	25.29Kb	25.29Kb	0.97	6.4Mb	0	Brute-force Linear crypt-analysis Differential crypt-analysis
		50kb	64	1.05	50.89Kb	50.89Kb	1.94	6.4Mb	0	
		75kb	64	1.26	76.39Kb	76.39Kb	2.87	6.4Mb	0	
2	3DES algorithm	25kb	168	1.391	25.3Kb	25.2Kb	0.869	6.7Mb	0	Linearization attack. chosen-plaintext known-plaintext
		50kb	168	1.698	50.9Kb	50.8Kb	1.484	6.9Mb	0.24	
		75kb	168	2.286	76.4Kb	76.3Kb	2.07	7.3Mb	0.24	
3	AES-128	25kb	128	0.48	25kb	25kb	0.25	7.0Mb	0	No known crypt-analytical attacks against AES but side channel attacks against AES implementations possible.
		50kb	128	0.56	50kb	50kb	0.46	7.2Mb	0	
		75kb	128	0.79	75kb	75kb	0.65	7.2Mb	0	
4	AES-192	25kb	192	0.67	25kb	25kb	0.67	7.2Mb	0	No known crypt-analytical attacks against AES but side channel attacks against AES implementations possible.
		50kb	192	0.95	50kb	50kb	0.7	7.2Mb	0	
		75kb	192	1.18	75kb	75kb	1.01	7.3 Mb	0	
5	AES-256	25kb	256	0.95	25kb	25kb	0.72	7.2Mb	0	No known crypt-analytical attacks against AES but side channel attacks against AES implementations possible.
		50kb	256	1.19	50kb	50kb	0.89	7.4Mb	0	
		75kb	256	1.543	75kb	75kb	0.94	7.8Mb	0	
6	IDEA	25kb	64	2.4	25Kb	25Kb	2.3	12Mb	17.3	Meet-in-the-middle attack. narrow- bicliques attack
		50kb	64	2.9	50Kb	50Kb	2.7	23Mb	20.4	
		75kb	64	3.2	75Kb	75Kb	3.1	33Mb	24	
7	Blowfish algorithm	25kb	64	1.29	25.2Kb	25.2Kb	1.12	12.5Mb	17.3	birthday attacks known-plaintext attacks
		50kb	64	2.5	50.8Kb	50.4Kb	2.2	24Mb	20.4	
		75kb	64	3.8	76.2.Kb	75.5Kb	3.37	33.7Mb	24	

Implementation is done for different symmetric encryption algorithms and three files sizes are considered as shown in table4. Comparison of these algorithms to find out the appropriate one for the proposed system. The following are the parameters:

1) size of plaintext
2) size of cipher text
3) size of decrypted text
4) key size
5) time taken to encryption plaintext
6) time taken to decryption of cipher text
7) CPU usage
8) Memory required
9) Vulnerability

From table 4, Comparing all the above parameters, Advance Encryption Standard (AES-128) is the best suitable for proposed system so the proposed work Advance Encryption Standard (AES) cryptographic algorithm for encryption of rules at user end.

The Rules based classifier database is given as input to AES-128 cryptographic algorithm for encryption, which produces 128-bit encrypted output and sent to server.

ISSN 1533-9211

DOI 10.5281/zenodo.7374949

At server side, the first step is to receive all decrypted database's sent by clients / counsellors.

Once all databases are decrypted, they are combined in one database. In combined database it is observed few rules are duplicated, few are so weak etc to resolve this problem optimization technique is introduced in proposed system.

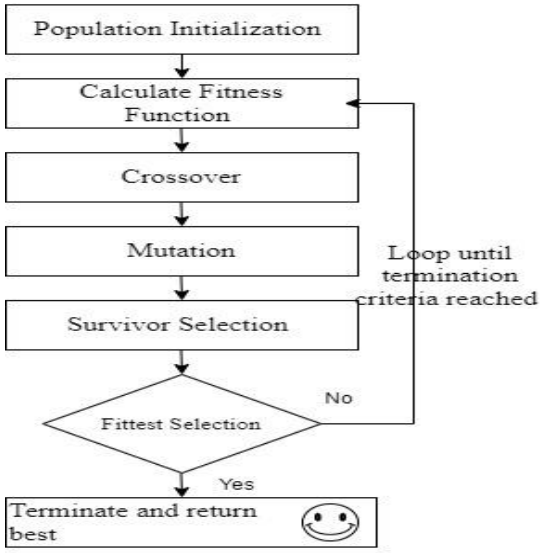
Part C: Optimization, is the method in which we train the model iteratively that effects in a maximum and minimum function value. It is one of the most vital phenomena in machine learning to get healthier results. Figure3 shows, the classification of optimization algorithms in machine learning. GA is evolutionary form intelligent optimization algorithm.

Genetic Algorithms are:

- They are robust in nature.
- Provide optimization over large space state.
- Unlike traditional AI, they do not break on slight change in input or presence of noise.
- Following is the study of different papers where genetic algorithm is used for optimization.

Research is present where GA is used for optimization ,in paper [28] author explained the effect of parameter fine-tuning with GA for solving traveling salesman problem. [29] Describe about population network construction effect on GA optimization performance. GA's are giving good results on Query optimization [30] as well as in Ontology Alignment optimization[31].

Figure 8: Basic structure of a GA



In GA's there is pond of possible solutions to the given problem. These solutions then undergo recombination and mutation, producing new children, the method is repeated over various generations. Each individual is assigned a fitness value. The fitter individuals are given a higher

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7374949

chance to mate and results into fitter individual. This way, keep evolving improved solutions over generations, till we reach a stopping criterion.

As the rules generated by the rule-based classifiers are mostly not mutually exclusive, i.e. many rules can cover the same record. The crucial goal of this research is usage of the genetic algorithm for optimization method. To reduce the number of duplicate and redundant rules and find the effective rules with the highest accuracy. Finally, there will be optimal and precise rules. By taking one or two rule from rule dataset, the step by step working is shown in figure 8.

Population Initialization:

- The Dataset on which we are working have 5000 instances and 27 attributes.
- Initial papulation is 100.

Calculate Fitness:

Rule	conf	cove	comp	fitness
(If Age = 35 & Mental health disorder in past = Yes & Family history = Yes & Seek help = Yes & No. of leaves = Yes then Positive detected)	0.79	0.70	1.0	0.73

Conf: confidence, Cove: coverage, Comp: Comprehensibility

Fitness: w1 * conf + w2 * cove + w3 * comp

w1, w2, w3 are wet selected by user and w1+w2+w3 =1, (w1=0.5, w2=0.3, w3=0.2)

Fitness = 0.5*0.79 + 0.3*0.70 + 0.2*1.0 = 0.735 = 73%

Accuracy = (TP + TN) / N = (255+147) / 500 = 80.4 %

Crossover:

(If Age = 35 & Mental health disorder in past = Yes & Family history = Yes & Seek help = Yes & No. of leaves = Yes then Positive detected)

(If Age = 25 & Gender =F & Treatment = Yes & Benefits =No & Care options =Not sure & Have mental health disorder = Maybe then Positive detected)

Chromosome genotype																											
Parent 1	18	0	0	1	0	0	0	0	0	0	0	0	2	0	3	0	0	0	0	0	0	0	2	0	0	0
Parent 2	9	2	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0

One-point crossover:

Parent 1	18	0	0	1	0	0	0	0	0	0	0	0	2	0	3	0	0	0	0	0	0	0	0	2	0	0	0
Parent 2	9	2	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0

Child 1	18	0	0	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	2	0	0	0
Child 2	9	2	0	1	0	0	0	0	0	1	1	0	0	0	3	0	0	0	0	0	0	0	0	0	2	0	0

ISSN 1533-9211

DOI 10.5281/zenodo.7374949

Mutation:

Child 1	**18**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**2**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**2**	**0**	**0**	**0**	**0**
Muted Child	**18**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**2**	**0**	**2**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**2**	**0**	**2**	**0**	**0**

Here two places are selected for mutation and their corresponding values get muted

Accuracy: (TP+TN) / N

N: total no of dataset instances

Rule:

If Age = 35 & Mental health disorder in past = Yes & Family history = Yes & Seek help = Yes & No. of leaves = Yes & Care option = Yes & Mental health disorder = Yes then Positive detected

Mutation takes places in such manner that rules become more specialized.

For finding Rule accuracy:

Accuracy = (TP + TN) / N = (203+243) / 500 = 89.2 %

- **Survivor Selection:**
  - **Fittest Selection:**

    Fitness: w1 * conf + w2 * cove + w3 * comp

    Conf: confidence

    Cove: coverage

    Comp: Comprehensibility

    w1, w2, w3 are wet selected by user and w1+w2+w3 =1, (w1=0.5, w2=0.3, w3=0.2)

Mined rule	conf	cove	comp	fitness
If age = 35 & mental health disorder in past = yes & family history = yes & seek help = yes & no. of leaves = yes & care option = yes & mental health disorder = yes then positive detected	0.97	0.89	1.0	0.95

Fitness: 0.5*0.97+0.3*0.89+0.2*1.0 = 0.952

Terminate and give best:

For above rule accuracy is 89.2 % and fitness is 0.95%

	Rule	Accuracy	Fitness
Rule before applying GA	(If age = 35 & mental health disorder inpast = yes & family history = yes & seek help = yes & no. of leaves = yes then positive detected)	**80.4 %**	**0.73**
Rule after applying GA	If age = 35 & mental health disorder inpast = yes & family history = yes & seek help = yes & no. of leaves = yes & care option = yes & mental health disorder = yes then positive detected	**89.2 %**	**0.95**

ISSN 1533-9211

DOI 10.5281/zenodo.7374949

Table 5: Comparison of rules before and after application of Genetic Algorithm

Sr. No.	Dataset	No of Attributes	No. of Instances	Before GA application the no. of Rules	After GA application the no. of Rules	Before GA application Best Accuracy (%)	After GA application Best Accuracy (%)
1	Mental Health Prediction	28	1257	20	16	65.8	96.5

Figure 9: Accuracy Comparison

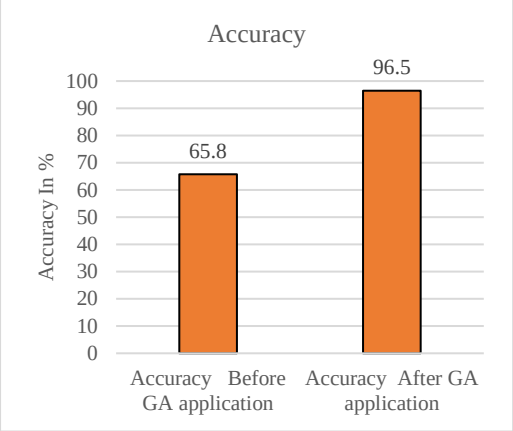


Figure10: Comparison of Rules

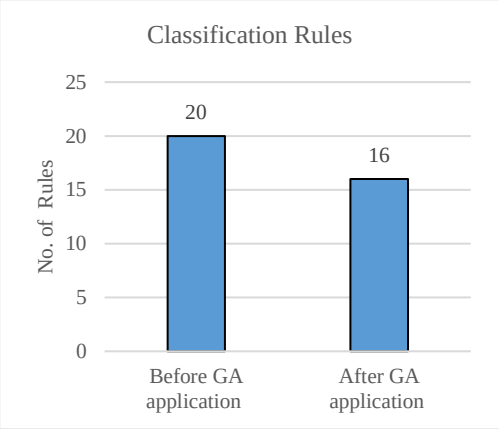


Table 5, shows the Comparison of rules before and after application of GA on Mental Health Prediction Dataset along with best accuracy as well figure 9, shows the best accuracy before and after application of genetic algorithm where accuracy before application of GA is 65.8 % which is increased up to 96 % after application of GA. Figure 10, shows the comparison of rules. The number of rules before application of GA is greater that the rules after application of GA, this is because the weak rules and duplicate rules are removed which is the aim of optimization.

The proposed algorithm is applied on different datasets below are the results.

ISSN 1533-9211

DOI 10.5281/zenodo.7374949

Table 6: Comparison of rules before and after application of GA on different datasets

Sr No.	Datasets	No of Attributes	No. of Instances	Before GA application the no. of Rules	After GA application the no. of Rules	Before GA application Best Accuracy (%)	After GA application Best Accuracy (%)
1	Mental Health prediction	27	5,000	428	219	77.262	99.84
2	Diabetes brf prediction	09	10,000	452	215	29.4	42
3	Stroke prediction	13	7,000	350	207	94	95.18
4	Thyroid prediction	11	5,110	546	304	93.74	95.75

Figure 11: Comparison of rules before and after application of GA

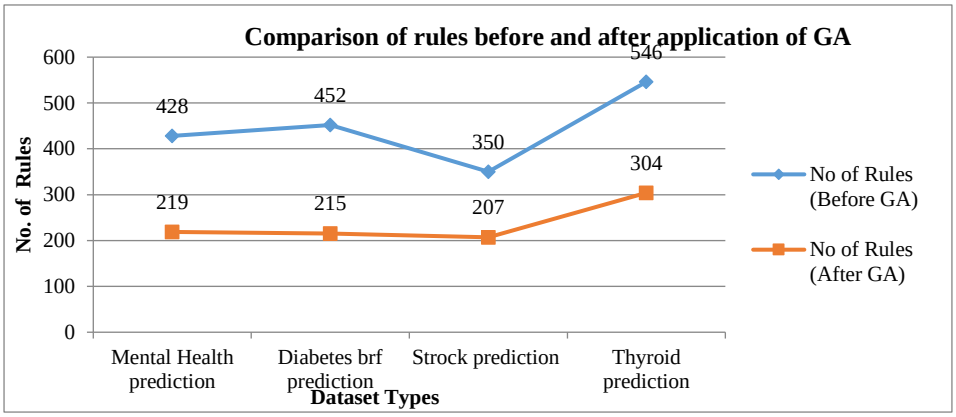


Figure 12: Rules accuracy comparison before and after application of Genetic algorithm

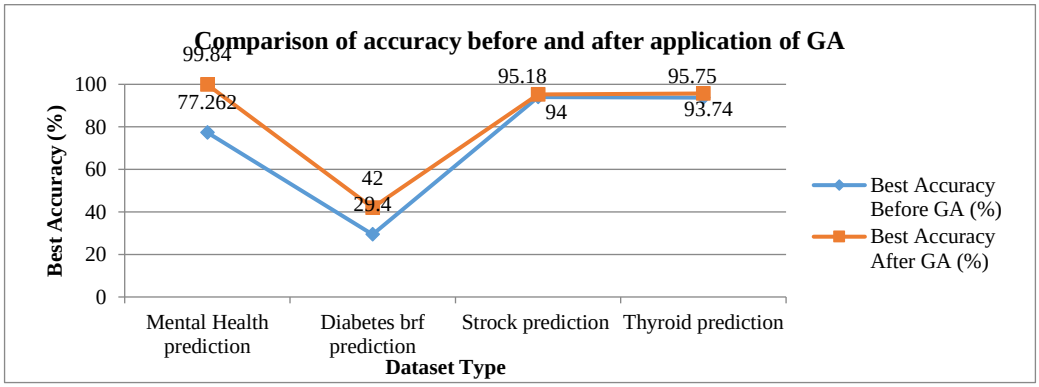


Table6, shows different datasets details like no. of instances, attribute with rules accuracy comparison before and after application of genetic algorithm on different datasets and before and after application of genetic algorithm the best accuracy. From table 6, it is observed that the proposed algorithm works better to increase the accuracy and to reduce the storage space,

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7374949

as no. of rules are optimized. Figure 11, shows comparison of rules before and after application of Genetic Algorithm, it is observed that the number of rules generated after application of GA are less that the rules generated by classifiers. Figure 12, shows the Rules accuracy comparison before and after application of Genetic Algorithm, where is observed the best accuracy of the rules is also increased after application of Genetic algorithm.

6. CONCLUSION

The proposed algorithm RaEnOpti works better to fix the problem of data sharing among hospital /counsellor with increase in the best accuracy. Which result in less diagnostic time as well less expenditure of different pathological test report for new cases. The proposed algorithm maintains the privacy of data by using PPDM's randomization techniques and secrecy by using symmetric encryption algorithm. While sharing the data storage space is very much important so by a using optimization technique that is also taken care. In nut shell the end user is getting are cleaned, optimized, and accurate data.

REFERENCES

[1] A. Abugabah, A. Al Smadi, and A. Abuqabbeh, "Data Mining in Health Care Sector: Literature Notes," PervasiveHealth Pervasive Comput. Technol. Healthc., pp. 63–68, 2019, doi: 10.1145/3372422.3372451.

[2] S. G. Alonso et al., "Data Mining Algorithms and Techniques in Mental Health: A Systematic Review," Journal of Medical Systems, vol. 42, no. 9. Springer New York LLC, Sep. 01, 2018, doi: 10.1007/s10916-018-1018-2.

[3] S. R. Agrawal Rakesh, "Privacy preserving data mining," Res. J. Appl. Sci. Eng. Technol., vol. 9, no. 8, pp. 616–621, 2015, doi: 10.19026/rjaset.9.1445.

[4] C. Lin, W. Lee, and Y. Ho, "Encryptions," in 19th International Conference on Advanced Information Networking and Applications (AINA'05) Volume 1 (AINA papers), Taipei, Taiwan, 2005, pp. 399-402 ,vol.2, doi: 10.1109/AINA.2005.99.

[5] C. C. Aggarwal and P. S. Yu, "Chapter 1 An Introduction to Privacy-Preserving Data Mining," pp. 1–2.

[6] L. Sweeney, "Generalizing Data to Provide Anonymity (Abstract) when Disclosing Information (Paper not available on time)," p. 188.

[7] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "ℓ-diversity: Privacy beyond k-anonymity," ACM Trans. Knowl. Discov. Data, vol. 1, no. 1, 2007, doi: 10.1145/1217299.1217302.

[8] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and-Diversity."

[9] X. Xiao and Y. Tao, "Personalized privacy preservation," Proc. ACM SIGMOD Int. Conf. Manag. Data, pp. 229–240, 2006, doi: 10.1145/1142473.1142500.

[10] V. Pinjarkar, A. Jain, A. Bhaskar, and P. Srivastava, "Pertinent Exploration of Privacy Preserving Perturbation Methods," Int. J. Recent Technol. Eng., vol. 8, no. 6, pp. 1945–1949, 2020, doi: 10.35940/ijrte.f8007.038620.

[11] R. Sanchez-Reillo, C. Lopez-Ongil, L. Entrena-Arrontes, and C. Sanchez-Avila, "Information technology security using cryptography," IEEE Aerosp. Electron. Syst. Mag., vol. 18, no. 6, pp. 21–24, 2003, doi: 10.1109/MAES.2003.1209586.

[12] Y. Nogami, "Pairing – based cryptography for homomorphic cryptography," in 2014 International

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7374949

Symposium on Information Theory and its Applications, Victoria, BC, Canada, 2014, pp. 318–321, [Online].

[13] P. Chaudhary, R. Gupta, A. Singh, and P. Majumder, "Analysis and Comparison of Various Fully Homomorphic Encryption Techniques," 2019 Int. Conf. Comput. Power Commun. Technol. GUCON 2019, pp. 58–62, 2019.

[14] S. Gupta and G. Arora, "Use of Homomorphic Encryption with GPS in Location Privacy," 2019 4th Int. Conf. Inf. Syst. Comput. Networks, ISCON 2019, pp. 42–45, 2019, doi: 10.1109/ISCON47742.2019.9036149.

[15] V. Pinjarkar and A. Jain, "Secured Data Transmission Using Different Techniques," in Tianjin Daxue Xuebao Journal of Tianjin University Science and Technology Vol.54, Issue 08, pp. 71–83, 2021, doi: 10.17605/OSF.IO/43YD8.

[16] S. C. Liang, Y. C. Lee, and P. C. Lee, "The application of ant colony optimization to the classification rule problem," Proc. - 2011 IEEE Int. Conf. Granul. Comput. GrC 2011, pp. 390–392, 2011, doi: 10.1109/GRC.2011.6122628.

[17] A. P. Engelbrecht and C. W. Cleghorn, "Recent advances in particle swarm optimization analysis and understanding 2021," GECCO 2021 Companion - Proc. 2021 Genet. Evol. Comput. Conf. Companion, pp. 341–368, 2021, doi: 10.1145/3449726.3461431.

[18] S. Chaturvedi, P. Pragya, and H. K. Verma, "Comparative analysis of particle swarm optimization, genetic algorithm and krill herd algorithm," IEEE Int. Conf. Comput. Commun. Control. IC4 2015, pp. 1–7, 2016, doi: 10.1109/IC4.2015.7375552.

[19] S. Wu, F. Zhang, and X. Wang, "Attribute Reduction Method Using Adaptive Genetic Algorithm and Particle Swarm Optimization," pp. 278–283, 2021.

[20] E. K. Naka and V. G. Guliashki, "Optimization Techniques in Data Management : A Survey," pp. 8–13.

[21] M. Saggar, A. K. Agrawal, and A. Lad, "Optimization of association rule mining using improved genetic algorithms," Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern., vol. 4, pp. 3725–3729, 2004, doi: 10.1109/ICSMC.2004.1400923.

[22] A. Hirabayashi, C. Aranha, and H. Iba, "Optimization of the trading rule in foreign exchange using genetic algorithms," Proc. IASTED Int. Conf. Adv. Comput. Sci. Eng. ACSE 2009, pp. 32–37, 2009.

[23] X. Zhang, H. Zhang, and E. Li, "Optimization of fuzzy classification system by genetic strategies," Proc. - 2010 6th Int. Conf. Nat. Comput. ICNC 2010, vol. 5, no. Icnc, pp. 2424–2428, 2010, doi: 10.1109/ICNC.2010.5583508.

[24] L. Yan, Z. Sun, and K. Mao, "Robust optimization based on an improved genetic algorithm," Adv. Mater. Res., vol. 655–657, no. 6, pp. 955–958, 2013, doi: 10.4028/www.scientific.net/AMR.655-657.955.

[25] M. Hassoon, M. S. Kouhi, M. Zomorodi-Moghadam, and M. Abdar, "Rule Optimization of Boosted C5.0 Classification Using Genetic Algorithm for Liver disease Prediction," 2017 Int. Conf. Comput. Appl. ICCA 2017, pp. 299–305, 2017, doi: 10.1109/COMAPP.2017.8079783.

[26] M. Interciso and P. B. Garcia, "Usage of a genetic algorithm for optimizing stock usage," GECCO 2020 Companion - Proc. 2020 Genet. Evol. Comput. Conf. Companion, vol. 3398159, no. 2, pp. 41–42, 2020, doi: 10.1145/3377929.3398159.

[27] B. Schneier and C. Systems, "Cryptography : The Importance Different," Computer (Long. Beach. Calif)., pp. 108–110.

[28] M. Mosayebi and M. Sodhi, "Tuning genetic algorithm parameters using design of experiments," GECCO 2020 Companion - Proc. 2020 Genet. Evol. Comput. Conf. Companion, pp. 1937–1944, 2020, doi: 10.1145/3377929.3398136.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.7374949

[29] A. Vié, “Population network structure impacts genetic algorithm optimisation performance,” GECCO 2021 Companion - Proc. 2021 Genet. Evol. Comput. Conf. Companion, pp. 1994–1997, 2021, doi: 10.1145/3449726.3463134.

[30] F. Li, “Research on Data Query Optimization based on Genetic Algorithm,” ACM Int. Conf. Proceeding Ser., no. 1, pp. 11–13, 2021, doi: 10.1145/3460179.3460182.

[31] X. Xue, X. Wu, and J. Chen, “Optimizing Ontology Alignment through an Interactive Compact Genetic Algorithm,” ACM Trans. Manag. Inf. Syst., vol. 12, no. 2, 2021, doi: 10.1145/3439772.

[32] V. Pinjarkar, A.Jain., "Analysis of Classification Rules Optimization Using Genetic Algorithm", in Journal of Design Engineering published at Issue-9, PP-4096-4107, December 2021.

[33] V. Pinjarkar, A. Jain., A. Bhasker, P Shrivastava , "Mental Health Disorder and Privacy Preserving Data Mining: A Survey", in book “The Role of IoT and Blockchain”, pp. 441–449, Taylor and Francis Group, (2022), 1 st Edition.

[34] V. Laijawala, A. Aachaliya, H. Jatta and V. Pinjarkar, "Classification Algorithms based Mental Health Prediction using Data Mining," 2020 5th International Conference on Communication and Electronics Systems (ICCES), 2020, pp. 1174-1178, doi: 10.1109/ICCES48766.2020.9137856.