# AN ENHANCED PRIVACY PRESERVING FRAMEWORK FOR DATA SANITIZATION AND RESTORATION



MD MOKHLESUR RAHMAN

UNIVERSITI KEBANGSAAN MALAYSIA

# AN ENHANCED PRIVACY PRESERVING FRAMEWORK FOR DATA SANITIZATION AND RESTORATION

MD MOKHLESUR RAHMAN



DISSERTATION SUBMITTED IN FULFILMENT FOR THE DEGREE OF MASTER OF INFORMATION TECHNOLOGY (COMPUTER SCIENCE)

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2022

# KERANGKA PEMELIHARAAN PRIVASI YANG DIPERTINGKAT UNTUK SANITISASI DAN PEMULIHAN DATA

MD MOKHLESUR RAHMAN



DISERTASI YANG DIKEMUKAKAN UNTUK MEMPEROLEH
IJAZAH SARJANA TEKNOLOGI MAKLUMAT (SAINS KOMPUTER)

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2022

**DECLARATION**

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

10 September 2022

MD MOKHLESUR RAHMAN
P99075

## ACKNOWLEDGEMENT

First and foremost, I would like to praise almighty Allah for His blessings for giving me power of mind, skill, patience and good health throughout the duration of this research work.

I would like to express my sincere gratitude and profound appreciation to my supervisor, Assoc Prof. Dr. Ravie Chandren Muniyandi, for his continuous guidance through advice, constructive criticisms and supports in every possible way throughout the completion of this research. I feel wonderful privileged for being able to work under his supervision for accomplishing this research work. Furthermore, I am very fortunate to have Assoc. Prof. Dr. Shahnorbanun Sahran and Dr. Suziyani Mohamed as my research co-supervisors who gave additional guidance and suggestions during this research.

I owe my success for this journey to my friends and classmates who made friendship and helped to create a nice working environment during the research work in the laboratory, UKM and also outside of the UKM. I also extend my gratitude to all academic personnel, staffs at FTSM for their supportive attitudes and UKM as well.

I gratefully acknowledge financial support provided by Universiti Kebangsaan Malaysia (UKM) and the Department of Higher Education, Malaysia Education Ministry, grant no. GGP-2019-023.

This thesis is wholeheartedly dedicated to my beloved parents, who prayed every single day and greatest source of inspiration. I am also lucky to have such parents-in-law who prayed and inspired for every steps. I would like to thank my sister, brother-in-law, my uncles and unties. Finally, it is not forgettable my dearest beloved wife Moslema @ Lipi Rahman who gave encouragements, motivations and supports when I was disappointed.

**ABSTRAK**

Gangguan Spektrum Autisme (ASD), gangguan perkembangan saraf, sering didedahkan pada zaman awal kanak-kanak. Penyelidikan ini menggunakan set data autisme bagi bayi berumur 24-, 30-, 36-, dan 48 bulan. Beberapa serangan biasa, seperti *spoofing, spamming, phishing*, serangan rangkaian baru, serangan berasaskan kandungan, serangan perkhidmatan teragih, serangan berasaskan perlombongan data dan ancaman pengenalan semula, mempengaruhi data ini. Terdapat juga beberapa serangan yang popular, contohnya, Serangan Sifir Diketahui (KCA), Serangan Teks Biasa Diketahui (KPA), Serangan Sifir Terpilih (CCA) dan Serangan Teks Biasa Terpilih (CPA). Jadi, privasi data adalah kebimbangan kritikal semasa memindahkan data untuk menghalang penjenayah siber daripada mengubah, mengganggu atau mencuri maklumat tersebut. Oleh itu, untuk melindungi data, penyelidik menggunakan pelbagai teknik, termasuk penyulitan cakera perisian atau perkakasan, pemadaman data, penutupan data, sandaran, dan pelbagai algoritma yang disulitkan dan dinyahsulit, seperti Standard Penyulitan Data (DES), Standard Penyulitan Lanjutan (AES), *Blowfish*, Algoritma Penyulitan Data Antarabangsa (IDEA), *Rivest Cipher 4* (RC4) dan lain-lain. Sebilangan kerja penyelidikan ini menggunakan k-tanpa nama dan pertanyaan, yang memerlukan sejumlah besar masa dan sumber pengiraan yang banyak. Selain itu, para penyelidik menggunakan algoritma pengoptimuman untuk memperbaiki isu privasi. Walau bagaimanapun, terdapat beberapa kekurangan, seperti tiada tempoh khusus untuk kemaskini nilai kunci semasa langkah penjanaan kunci, tidak menyebut nilai panjang kunci, tidak mentakrifkan nilai parameter, bilangan julat kunci yang tidak ditentukan, dan ketidaksesuaian fungsian *fitness*. Untuk menangani keurangan ini, kajian mencadangkan kerangka algoritma meta-heuristik yang dipanggil Kerangka Gabungan PSO-GWO Dipertingkatkan (Pengoptimuman Kawanan Zarah-Pengoptimuman Serigala Kelabu) (*Particle Swarm Optimization-Gray Wolf Optimization*). Kerangka ini menggunakan dua teknik, iaitu, prosedur sanitasi data dan pemulihan data. Pada mulanya, kajian mencipta kunci optimum, yang digunakan dalam proses sanitasi data. Seterusnya, kunci yang sama digunakan dalam proses pemulihan untuk memulihkan data. Kajian ini membandingkan prestasi kerangka yang dicadangkan dengan algoritma tradisional, seperti PSO (*Particle Swarm Optimization*), GA (*Genetic Algorithm*), DE (*Differential Evolution*), CSA (*Crow Search Algorithm*) dan AAP-CSA (*Adaptive Awareness Probability* berasaskan CSA) terhadap serangan popular yang disebutkan di atas dan mencapai prestasi yang lebih baik. Daripada simulasi data sanitasi, didapati teknik yang dicadangkan, dari segi serangan KPA, mencapai 99.87%, 99.77%, 99.47%, 99.26%, dan 99.72%, iaitu lebih ketara bertambah baik berbanding PSO, GA, DE, CSA, dan AAP-CSA, untuk tempoh 30 bulan bagi set data autisme bagi kebanyakan jenis data autisme yang lain. Sebaliknya, untuk pemulihan data, model menunjukkan daripada simulasi bahawa ia mencapai 99.89%, 99.81%, 99.54%, 99.37% dan 99.76%, yang dipertingkatkan berbanding PSO, GA, DE, CSA dan AAP-CSA, masing-masing, di bawah set data kanak-kanak autisme 30 bulan, bagi kebanyakan antara jenis data autisme yang lain.

**ABSTRACT**

Autism Spectrum Disorder (ASD), a neurodevelopmental disorder, is often unveiled in early childhood. This research utilized autism datasets of 24-, 30-, 36-, and 48-months old babies. Some common attacks, such as spoofing, spamming, phishing, novel network attacks, content-based attacks, distributed service attacks, data mining-based attacks, and re-identification threats, affect these data. There are also some popular attacks, for example, the Known Cipher Attack (KCA), Known Plaintext Attack (KPA), Chosen Cipher Attack (CCA), and Chosen Plaintext Attack (CPA). So, data privacy is a critical concern while transferring data to prevent cyber criminals from altering, interrupting, or stealing the information. Consequently, to protect data, researchers employ a variety of techniques, including software or hardware disk encryption, data erasure, data masking, backup, and various encrypted and decrypted algorithms, such as Data Encryption Standard (DES), Advanced Encryption Standard (AES), Blowfish, International Data Encryption Algorithm (IDEA), Rivest Cipher 4 (RC4), and others. A number of these research works make use *k*-anonymity and query, which need a significant amount of time and substantial computational resources. Moreover, the researchers are employing optimization algorithms to improve the privacy issue. However, there are some limitations, such as no specific duration for updating the key value during the key generation step, not mentioning the key length based on which value, not defining the values of the parameters, an undefined number of key ranges and inappropriate fitness functions. To address these critical and significant concerns, this research proposed a meta-heuristic algorithmic framework called the *Enhanced Combined PSO-GWO* (*Particle Swarm Optimization-Grey Wolf Optimization*) *Framework*. This framework employed two techniques, which are data sanitization and data restoration procedures. Initially, the study creates optimal key, which is employed in the data sanitization process. After that, the same key is employed in the restoration process also to restore the data. This study compared the performances of the proposed framework with the traditional algorithms, such as PSO (Particle Swarm Optimization), GA (Genetic Algorithm), DE (Differential Evolution), CSA (Crow Search Algorithm), and AAP-CSA (Adaptive Awareness Probability-based CSA) against the above-mentioned popular attacks and achieved better performances. From the simulation for sanitizing data, it is revealed that the proposed technique, in terms of KPA attack, attained 99.87%, 99.77%, 99.47%, 99.26%, and 99.72%, which are more significantly improved over PSO, GA, DE, CSA, and AAP-CSA, respectively, over the 30 months autism dataset, mostly among the other types of autism data. On the other hand, for restoring data, the model shows from the simulation that it achieved 99.89%, 99.81%, 99.54%, 99.37%, and 99.76%, which are enhanced over PSO, GA, DE, CSA, and AAP-CSA, respectively, under the 30 months autism child dataset, mostly among the other types of autism data.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF ILLUSTRATIONS

## LIST OF ALGORITHMS

**LIST OF ABBREVIATIONS**

| | |
|---|---|
| AA | Anonymous Authentication |
| AAP-CSA | Adaptive Awareness Probability-based Crow Search Algorithm |
| ABC | Artificial Bee Colony |
| ABE | Attribute-Based Encryption |
| ABIDE | Autism Brain Imaging Data Exchange |
| ACM | Access Control Manager |
| ADHD | Attention Deficit Hyperactivity Disorder |
| ADI-R | Autism Diagnostic Interview – Revised |
| ADOS | Autism Diagnostic Observation Schedule |
| AES | Advanced Encryption Standard |
| AGRE | Autism Genetic Resource Exchange |
| AMP | Autism Management Platform |
| ANN | Artificial Neural Network |
| ASD | Autism Spectrum Disorder |
| ASQ:SE (M) | Ages & Stages Questionnaires: Social-Emotional (Malaysia) |
| BS-WOA | Brain Storm-based Whale Optimization |
| CCA | Chosen Cipher Attack |
| CI-LA | Crossover Improved-Lion Algorithm |
| CloudDLP | Cloud Data Loss Prevention |
| CPA | Chosen Plaintext Attack |
| CP-ABE | Ciphertext-Policy Attribute-Based Encryption |
| CSA | Crow Search Algorithm |
| CSA | Cuckoo Search Algorithm |
| DDoS | Distributed Denial of Service |
| DE | Differential Evolution |
| DES | Data Encryption Standard |
| DLP | Data Loss Prevention |

| | |
|---|---|
| DM | Degree of Modification |
| DSM | Diagnostic and Statistical Manual of Mental Disorders |
| ECDH | Elliptic Curve Diffie-Hellman |
| eHF | e-Health Framework |
| EPR | Electronic Patient Record |
| FBI | Federal Bureau of Investigation |
| FR | False Rule |
| GA | Genetic Algorithm |
| GBPSO-SVM | Geometric Binary Particle Swarm Optimization-Support Vector Machine |
| GEO | Gene Expression Omnibus |
| GMGW | Genetically Modified Glowworm Swarm Optimization |
| GPIT | Global Patient Identification Technique |
| GSA | Gravitational Search Algorithm |
| GSO | Glowworm Swarm Optimization |
| GWO | Grey Wolf Optimization |
| HF | Hiding Failure |
| IBE | Identity-Based Encryption |
| IDEA | International Data Encryption Algorithm |
| IEEE | Institute of Electrical and Electronics Engineers |
| IMSICF | Improved Maximum Sensitive Itemsets Conflict First |
| IoMT | Internet of Medical Things |
| IP | Information Preservation |
| JA | Jaya Algorithm |
| J-SSO | Jaya-based Shark Smell Optimization |
| KCA | Known Cipher Attack |
| KPA | Known Plaintext Attack |
| LA | Lion Algorithm |
| LDA | Linear Discriminant Analysis |

| | |
|---|---|
| LSB | Least Significant Bit |
| ML | Machine Learning |
| MRI | Magnetic Resonance Imaging |
| NCBI | National Center for Biotechnology Information |
| NDAR | National Database for Autism Research |
| NN | Neural Network |
| NRGR | National Institute of Mental Health Repository and Genomics Resource |
| OI-CSA | Opposition Intensity-based Cuckoo Search Algorithm |
| PASH | Privacy-Aware Smart Health |
| PGVIR | Parallelized Grouped Victim Item Removal |
| PHCR | Parallelized Hiding Candidate Removal |
| PHDMS | Personal Health Data Management System |
| PKE | Public Key Encryption |
| PPUM | Privacy-Preserving Utility Mining |
| PSO | Particle Swarm Optimization |
| PSO-GWO | Particle Swarm Optimization- Grey Wolf Optimization |
| RC4 | Rivest Cipher 4 |
| ROI | Region Of Interest |
| RONI | Region Of Non-Interest |
| rs-fMRI | resting state-functional Magnetic Resonance Imaging |
| SE | Searchable Encryption |
| SHRs | S-Health Records |
| Simons VIP | Simons Variation in Individuals Project |
| SKE | Symmetric Key Encryption |
| sMRI | Structural Magnetic Resonance Imaging |
| SRS | Social Responsiveness Scale |
| SSC | Simplex Simon Collection |
| SSE | Searchable Symmetric Encryption |

| | |
|---|---|
| SSO | Shark Smell Optimization |
| SVC | Support Vector Classification |
| SVM | Support Vector Machine |
| SVM-RFECV | Support Vector Machine-Recursive Feature Elimination with a stratified-4-fold Cross-Validation |
| TD | Typical Development |
| TripleDES | Triple Data Encryption Standard |
| UCI | University of California Irvine Machine Learning Repository |
| UKM | Universiti Kebangsaan Malaysia |
| USA | United States of America |
| WEKA | Waikato Environment for Knowledge Analysis |
| W-iPCN | Wireless Intelligent Personal Communication Node |
| WNU | Whale with New Crosspoint-based Update |
| WOA | Whale Optimization Algorithm |

# CHAPTER I

# INTRODUCTION

## 1.1 RESEARCH BACKGROUND

With the advancement of technology, a large amount of data in electronic form is being gathered. One of the major sectors is medical applications and medical equipment, from where medical health data is produced and transferred over the networks for improvement of the health sector (Mewada et al. 2020). The crucial issue in the transmission of this data is the preservation of its privacy (Zaabar et al. 2021). As a medical dataset, autism data has been selected to be employed for this research.

Autism is a complex neurobehavioral condition that starts with the onset of early childhood. It impairs social interactions and communication skills combined with restricted, repetitive behaviours according to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) (American Psychiatric Association 2013; Thabtah 2017). Because of the range of symptoms, this condition is now called autism spectrum disorder (ASD). Because of having a distinct set of deficits, DSM-5 has divided ASD into three levels of severity based on the amount of support required by children with ASD in their daily lives, such as requiring support (Level 1), requiring substantial support (Level 2), and requiring very substantial support (Level 3), respectively. Children with ASD demonstrate poor social-communication skills as they have deficits in verbal communication, non-verbal communication, and social-emotional reciprocity. Deficits are, in this case, nonunderstanding of spoken language, inappropriate tone of voice during conversation; nonunderstanding the meaning of body gestures, inappropriate facial expressions; difficulties in expressing their own emotions, and recognizing their own and others' emotions, etc. Children with ASD also demonstrate restricted and repetitive patterns of behaviour, interests, or activities. Deficits in this

area are characterized by repetitive body movements and motions, ritualistic behaviours, and a restricted or extreme interest in specific activities or objects. However, the transmission of this autism data over the network needs to maintain privacy and protection, which is more essential recently because a large number of data are congregating every time (Sivan & Zukarnain 2021).

The data concerning health is known as sensitive data that requires additional protection since it may get into the very core of a human being. As a result, the main concentration of this thesis is the data privacy regarding autism sensitive data. There are some critical concerns and challenges that are very significant for the privacy of healthcare data, such as i) unauthorized access, ii) data breach, iii) data disclosure, iv) data modification, v) data forgery, and so on (Oh et al. 2021):

i. Unauthorized access: Unauthorized access is a possibility due to a number of safety flaws in medical equipment, networks, and platforms, including edge, fog, and cloud. An attacker can gain access to the system and take control of sensitive e-health data by taking advantage of their weaknesses.

ii. Data breach: Data breach is a protection breach wherein private, protected, or sensitive information is duplicated, transferred, accessed, or utilized by someone not authorized. It may comprise personal health information, financial details including credit card or debit card numbers, bank account information, trade secrets of business or property, etc.

iii. Data disclosure: Due to administrative error or safety flaws in medical equipment, networks, and cloud platforms, data disclosure may take place across the e-healthcare system. Due to the high value of e-health data, attackers are drawn to it.

iv. Data modification: This is the act of altering data without the necessary authentication and authorisation. This assault, also known as data tampering, may pose a serious security risk since altered e-health data might have significant consequences for patients.

v. Data forgery: E-health information or user IDs can be falsified, known as data forgery, to mislead reliable service providers or pass as someone else.

E-health systems can be compromised by an attacker using fake information, or a user with malevolent intent can get profit improperly. So, in order to protect patient privacy, e-health data should be anonymized when it needs to be shared or public. Specifically, it is necessary to anonymize e-health data which is irrelevant to a specific reason and identify data on patients and healthcare personnel that may be used to connect anonymized data to individuals. This data anonymization is also known as data sanitization which has the goal of preserving privacy. This is a method of delating personally identifying data from datasets in order to keep data anonymous against the unauthorized personnel (Sei et al. 2019). There are various techniques and methods for anonymization of data, such as k-anonymity, l-diversity, t-closeness, differential privacy, etc., that are employed in several research. Though k-anonymity, proposed by Sweeney (Sweeney 2002), is better for the individual not to be identified, but it depends on the optimal k value. So, the concern under the k-anonymity is the likelihood of re-identification increases with more similar the sensitive features are. As a result, l-diversity is suggested by Machanavajjhala et al. (Machanavajjhala et al. 2006) that can defend the drawbacks of k-anonymity by doing diversify the sensitive traits. Similar to k-anonymity, data privacy is improved by increasing the l value according to l-diversity. But l-diversity has also limitations that privacy can be revealed if the anonymized data contain biases or patterns. So, to mitigate the limitations of k-anonymity and l-diversity, t-closeness was recommended by Li et al (N. Li et al. 2007). Moreover, if the attackers have prior knowledge of manipulating vulnerabilities, the data privacy for k-anonymity and l-diversity can be hampered. Dwork et al. (Dwork et al. 2006) tried to overcome this problem by offering differential privacy. By introducing noise into every single query, this approach stops an attacker from identifying a particular person from statistical data obtained from several database queries. The attacker is unable to determine the distribution of data that may be utilised for re-identification because of this noise. In the cryptosystem, advanced encryption standard (AES) and Rivest–Shamir–Adleman (RSA) are utilized in purpose of the e-health data security where symmetric key was used for fast encryption and decryption speed in AES as well as a public key for encryption and a private key for decryption are utilized in RSA.

Moreover, there are various optimization algorithms, such as whale optimization algorithm (WOA), artificial bee colony (ABC), maximum sensitive

itemset conflict first algorithm (MSICF), crow search algorithm (CSA), particle swarm optimization (PSO), grey wolf optimization (GWO), genetic algorithm (GA), differential evolution (DE), adaptive awareness probability-based CSA (AAP-CSA), etc., that are also being utilized for the purpose of data privacy in the abundant research (Abidi et al. 2021; Lekshmy & Rahiman 2019; X. Liu et al. 2020; Mewada et al. 2020). Most of the studies yielded appropriate keys using these algorithms for the privacy of medical data (Alphonsa & Amudhavalli 2018; Mandala & Rao 2019; Mewada et al. 2020). In this case, some significant issues need to be addressed, such as specifying the duration for updating the key values during the key generation step, mentioning the key length, defining appropriate parameters, defining key ranges, etc. In addition, measuring the appropriate objective functions through the respective algorithms in the sanitization and restoration processes to sanitize and restore the data in order to protect the privacy of data.

Therefore, yielding the optimal key is an important privacy issue with the appropriate optimization algorithms. In this thesis, a framework has been proposed where the suitable optimal key is produced by combining two optimization algorithms, such as particle swarm optimization (PSO) and grey wolf optimization (GWO). After that, the key is employed in the sanitization process and restoration process, and the appropriate objective functions for those processes are measured for the purpose of ensuring the privacy of autism data.



## 1.2 MOTIVATION

Data privacy implies to the dealing with the capability of an individual or an organization that determine what types of data into the transmission line should be distributed with each other. Whereas data protection indicates to the defensive digital safety procedures that are utilized to halt unlawful control over databases. There are manifold ways for data privacy and protection, including software or hardware disc encryption, data erasure, data masking, backup, and advanced technologies, as examples of data protection. However, data privacy is considered as an important issue of information sharing. Sharing information in a way that defends personally identifiable data which is a key challenge in the data privacy. Medical data, for example,

autism datasets are highly private (Khan & Hoque 2016). This information pertains to a person's most personal life. Unauthorized disclosure can result in a variety of types of discrimination and violations of basic rights (Seh et al. 2020; Vezyridis & Timmons 2019). For example, it could be exploited, misused, or misinterpreted for a certain purpose (Coventry & Branley 2018; Eling & Wirfs 2019; Volkmar et al. 2012). It also might have an impact on someone's insurance or job. Individuals or organisations who process and keep medical information are frequently compelled to use privacy and safety methods. In this regard, many countries have created confidentiality-preserving standards for doctor-patient relationships that preserve confidentiality (Raspa et al. 2020). These standards preserve patients' dignity while also ensuring that they provide accurate information in order to obtain the best treatment possible.

The healthcare data is extremely vulnerable because of cyber hackers (Shah & Khan 2020). This data is increasingly being hacked nowadays because hackers' goal is to misuse personal data, which is a profitable financial business for them (Chernyshev et al. 2019; Coventry & Branley 2018; Sulleyman 2017; Trustwave 2017). Such as names, dates of birth, health insurance numbers, diagnostic codes, and billing information, etc., are among the medical data available to sell for business purposes. Fraudsters utilise this information to open bank account, get passport (Kangas 2017) or construct forged identification cards in order to manufacture medical equipment or medications. Moreover, according to specialists who have researched cyber-attacks on healthcare institutions, hackers combine a patient number with a bogus provider number, create a file, and then submit a claim with the insurance companies (Khan & Hoque 2016). One of the estimations of FBI is that $80 billion of the $2.2 trillion a year have been spent on healthcare in the USA, which is related with fraud, with half of that fraud tied to medical ID theft (McGee 2014). Since hospitals or clinics have interconnection, easily accessible access points, outdated systems, and a lack of emphasis on cybersecurity (Coventry & Branley 2018) for sharing data, hackers may easily obtain enormous amounts of data. Therefore, data privacy of e-health data is very significant to emphasize (Boonyarattaphan et al. 2009; Chernyshev et al. 2019; Price & Cohen 2019; Sivan & Zukarnain 2021) that is one of the major concerns nowadays.

## 1.3 RESEARCH PROBLEMS

There are various frameworks or models with different techniques, methods or algorithms which maintain the data privacy issue. Among them, there have numerous models, such as Model-Driven Application-Level Encryption for the Privacy of E-Health Data, Security Model for Big Healthcare Data Lifecycle, Global Patient Identification Technique (GPIT), Autism Management Platform (AMP), Attribute-Centric Anonymization Scheme, Adaptive Awareness Probability-based Crow Search Algorithm (AAP-CSA), Personal Health Data Management System (PHDMS), and etc. (Ding & Klein 2010; Khaloufi et al. 2018; Khan & Hoque 2016; Linstead et al. 2016; Majeed 2019; Mandala & Rao 2019; Wu et al. 2016) which tried to maintain data privacy while transferring medical health data. Those frameworks or models utilized various algorithms or techniques, for instance, grey wolf optimizer algorithm, authentication, data encryption, data masking, access control, monitoring and auditing, d-identification, hybrid execution, identity based anonymization, larger keys, particle swarm optimization, encryption/decryption, number variance, shuffling records, substitution, key encryption, imputation methods, data augmentation, genetic algorithm, differential evaluation, and crow search algorithm. (Abdel-Basset et al. 2020; Abouelmehdi et al. 2018; Al-Tashi et al. 2020; Bernstein 2005; Bonyadi & Michalewicz 2017; Edgar 2004; Mirjalili et al. 2020; Sivan & Zukarnain 2021; Thomas et al. 2020; Zolghadr-Asli et al. 2018). However, those frameworks or models have still possessed some lacking while they are using those algorithms or techniques in purpose of data privacy. The lacking are regarding for generating optimal key purposes, for instance, how long the key value will be updated during the key generation stage, the key length will be allocated based on which value, parameter setting, number of the key ranges; not measuring appropriate performance matrices which have significant effects on sanitizing and restoring data (Mewada et al. 2020); and convergence issues, like convergence to a point (known as stability), slower convergence, not enough local searching ability, etc. of optimization algorithms, such as Particle Swarm Optimization (PSO) algorithm (Bonyadi & Michalewicz 2017), Grey Wolf Optimizer (GWO) algorithm (Al-Tashi et al. 2020; Mirjalili et al. 2020).

PSO is a nature-inspired evolutionary robust stochastic optimization approach for solving computationally difficult problems. A swarm of *n* particles (individuals), in the PSO, interacts with each other via search direction (gradients) either directly or indirectly. It is made up of three vectors: the x-vector maintains the particle's present position (location) in the search space, the p-vector (pbest) represents the location of the particle's best solution thus, and the v-vector holds the gradient (direction) in which the particle will visit if undisturbed. In addition, Grey Wolf Optimizer (GWO), is a swarm intelligent technique which resembles the leadership hierarchy of wolves who are widely recognised for their collective hunting approach. The hierarchy is carried out by classifying the search agent population into four sorts of individuals depending on their fitness: level 1 (Alpha), level 2 (Beta), level 3 (Delta), level 4 (Omega). In addition, these two meta-heuristic algorithms are illustrated in section 2.4 (Traditional Meta-heuristic Algorithms) broadly and also revealed as how to be utilized in improving data privacy in Chapter IV and Chapter V. To address the above problems, the thesis has incorporated the characteristics of PSO into the GWO to enhance the capability of convergence as well as local searching ability whose main purpose is to generate optimal key for the purpose of sanitization and restore the autism sensitive data. In the proposed framework, the performance matrices that are introduced are information hiding failure rate, information loss rate and degree of modification rate. Therefore, the problem statements are summarized in the following:

i. Generating optimal key is a significant problem to sanitize sensitive data in order to attain privacy autism data.
ii. Defining perfect parameters and measuring appropriate performance matrices or objective functions for two processes, such as sanitization process and restoration process, are important issues for data privacy of medical data.
iii. Enhancing the search ability of different optimization algorithms for solving optimization problems in a framework is a big challenge that plays an important role in obtaining the best solution.

The purpose of this thesis is to form a framework that uses state-of-the-art technologies for data privacy. Specifically, this work applied data sanitization and data

restoration techniques by creating optimal key for privacy of data. Here, the work emphasized the privacy issue regarding autism datasets.

## 1.4 RESEARCH QUESTIONS

This study poses some research questions, which are mentioned as follows:

1. What are the existing techniques and suitable methods available that provide data privacy and identifying limitations and strengths?
2. How the sanitization and restoration techniques can be used to improve optimization and privacy?
3. How can a framework be created that would be better for data privacy in autism?

## 1.5 OBJECTIVES OF THIS RESEARCH

The objective of this study is to discuss the current state of autism data privacy and develop a new framework that provides the privacy of autism datasets. However, this research intends to achieve the following objectives:

1. To generate the optimization key and enhance sanitization process to sanitize data for ensuring the privacy of the autism datasets.
2. To enhance a restoration process to restore data for ensuring the privacy of the autism datasets.
3. To design a framework that maintains the privacy of autism datasets.

## 1.6 SCOPE OF THE RESEARCH

The scope of this thesis is to cover the privacy issue of medical data, such as autism sensitive data. For this reason, this study creates a framework which employs two techniques, namely sanitizing technique and restoration technique and utilizes two meta-heuristics algorithms, particle swarm optimization (PSO) and grey wolf optimization (GWO), combinedly. An optimal key is obtained at first by this framework

and is applied into two processes, such as sanitization process to sanitize the medical data as well as in restoration process to restore those data. In these two processes, two individual algorithms are employed for sanitization and restoration purposes. The performance of this framework is compared with some existing conventional algorithms, such as genetic algorithm (GA), particle swarm optimization (PSO), crow search algorithm (CSA), differential evolution (DE), and adaptive awareness probability-based CSA (AAP-CSA) against some popular attacks, for example, known cipher attack (KCA), known plaintext attack (KPA), chosen cipher attack (CCA), and chosen plaintext attack (CPA). This research used four different types of autism datasets like autism child dataset 24 months, autism child dataset 30 months, autism child dataset 36 months, and autism child dataset 48 months. These data have been pre-processed using support vector machine (SVM). All the datasets were collected from the Centre of Community Well-being and Education, Faculty of Education, Universiti Kebangsaan Malaysia.

## 1.7 ORGANIZATION OF THE THESIS

This thesis works with the enhancement of the privacy of autism sensitive datasets. The arrangement of this thesis is structured as below:

The research work consists of six chapters including different sections and subsections. In the beginning, I introduce the Chapter I as Introduction with the research background, motivation, research problems, aims of this research, research questions, specific objectives, and the scope of this research work.

The research has discussed and analysed the contributions, the research gabs of the previous research works regarding privacy of data, especially for autism data in Chapter II. A summary of existing techniques or algorithms in different frameworks utilized by the researchers is presented also. There are two meta-heuristics algorithms, namely particle swarm optimization (PSO) and grey wolf optimization (GWO), applied in the recommended framework has been illustrated as well. Moreover, the very specific descriptions and research gaps in the existing literatures regarding optimization key, sanitization and restoration for autism sensitive data, and their respective summarization tables are provided separately in this chapter.

The research methodology is demonstrated in Chapter III. After starting with the introduction section, this chapter introduce the research phases during this research work. After that, in section 3 and section 4, a broad discussion on the methodology for this thesis is illustrated through the major three key points for obtaining the specific objectives. Section 5 describes the four autism child datasets in detail which are used in this research. A brief discussion on WEKA Software and Python programming language have been discussed in afterwards section. Finally, section 8 concludes the methodology chapter.

Chapter IV demonstrates the key generation, key extraction, key encoding, key transformation, etc. to get the optimal key and sanitization process for sanitizing autism data for privacy. After that, the study has discussed how those two specific meta-heuristic algorithms work in the recommended Enhanced Combined PSO-GWO framework. The simulation and discussions section simulates and reveals the enhancement of the proposed technique by comparing with some existing algorithms in terms of different kinds of attacks on the four autism child datasets.

On the other hand, restoration procedures of autism sensitive data by the proposed Enhanced Combined PSO-GWO are discussed in Chapter V broadly. Here, how those two meta-heuristic algorithms work in the framework for restoring data is illustrated in detail. The simulations are performed on different values of acceleration constants against cost function and achieves the contributions of the proposed framework in comparison with existing conventional algorithms. The four types of autism datasets are utilized also in this simulation.

Finally, the thesis is concluded in Chapter VI, where a summary of investigations, findings, and contributions of the research is discussed. Lastly, the future works are recommended that need to be addressed.

# CHAPTER II

# LITERATURE REVIEW

## 2.1 INTRODUCTION

Data privacy is the fundamental concern for data sharing through the internet. It implies dealing with the capability of a person or an organization to control which types of data through the transmission line should be exchanged, among others. Therefore, it is considered an important issue of information sharing. Medical data, especially among other data, is a highly lucrative business for cyber hackers. So, medical data is under the threat of being stolen by cyber criminals (Omar et al. 2021). As a case study, this research has taken the autism dataset into consideration because of its importance.

An increasing number of children identified with ASD is occurring worldwide day by day (Maenner et al. 2020). Accordingly, the amount of autism data is increasing, which supports the importance of early assessment and treatment using evidence-based treatments, which may greatly enhance the quality of life for people with ASD, their caregivers, and their families (Elder et al. 2017). As a result, the growing amount of autism data should be cause for concern, as this data can be leaked while being transferred across the network. However, the critical issue raised by the transmission of data is the privacy issue (Zaabar et al. 2021).

The privacy of data is essential because many medical datasets are being gathered (Mandala & Rao 2019). In this work, the privacy of data has been emphasized because of its importance (Mewada et al. 2020; Sivan & Zukarnain 2021; W. Wang et al. 2015). To maintain privacy and protection, various frameworks or models use different techniques, methods, or algorithms that attain data privacy and protection issues, such as cost-effective and model-driven application-level frameworks for e-

health data transmission using different encrypted and decrypted algorithms, namely, DES, 3DES or TripleDES, AES, Blowfish, IDEA, and RC4. Researchers are trying to minimise the increasing interruption of data flow or threats of hacking by utilizing various advanced techniques (Alesawy & Muniyandi 2016; He et al. 2017; Idowu & Muniyandi 2019; Sowjanya et al. 2021; Zaman et al. 2017). Some researchers also utilize various meta-heuristics algorithms like artificial bee colony (Mewada et al. 2020), crow search algorithm (CSA) (Zolghadr-Asli et al. 2018), particle swarm optimization (PSO) (W. Li et al. 2021; D. Wang et al. 2018), glowworm swarm optimization (GSO) (Alphonsa & Amudhavalli 2018), grey wolf optimizer (GWO) (Y. Li et al. 2021; Panda & Das 2019), and so on. To address the privacy and protection problems, some of these investigations use k-anonymity and query. Such approaches need a large amount of time and computer resources. Besides, some of these traditional meta-heuristics algorithms also possess lower solving precision, slower convergence, and worse local searching ability. Some of them employed data sanitization process using optimized key (Lekshmy & Rahiman 2019; X. Liu et al. 2020). Some important issues need to be tackled in the case of forming optimal keys, such as indicating the duration for updating the key values during the key generation step, referring to the key length, specifying appropriate parameters, setting the key ranges, etc. Data sanitization is a technique where data is hidden by using appropriate keys. Moreover, a very limited number of articles worked against privacy regarding autism data using restoration procedure by privacy preservation models (Abidi et al. 2021; Ahamad et al. 2022; Alphonsa & Amudhavalli 2018; Mandala & Rao 2019; Mewada et al. 2020; Shailaja & Rao 2020), though the accuracy of these models seems to be inadequate. Data restoration is the process of copying backup data from secondary storage and restoring it to its original location or a new location (Balashunmugaraja & Ganeshbabu 2020). The data restoration is important to ensure the efficient restoration of the real data and to built-up cyber. To enhance the performances of sanitization and restoration procedures, there is a need to improve the processes (Ahamad et al. 2022) as well as the objective functions (Mewada et al. 2020).

So, for the privacy of data, it is necessary to address the above issues. A large amount of medical data is now accumulating in health data storage. Therefore,

researchers are becoming more conscious about the privacy of data to develop their research work for contributing to the world.

## 2.2 OVERVIEW OF AUTISM DATA

In this section, various sources of autism data, imbalanced data and their tackling processes, as well as features along with their types of sample autism data are illustrated below:

### 2.2.1 Sources of Autism Dataset

Most of the researchers utilize the various repository for autism dataset in their research works. There are different sites where autism related datasets can be collected for research purposes. Table 2.1 shows various primary sources of autism datasets. Researchers do their research works on these different types of autism datasets most of the cases.

Table 2.1 Various sources of autism datasets

| SL No. | Data Source | Website |
|---|---|---|
| 01 | UCI | https://archive.ics.uci.edu/ml/datasets.php |
| 02 | ABIDE I | http://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html |
| 03 | ABIDE II | http://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html |
| 04 | NDAR | https://ndar.nih.gov |
| 05 | AGRE | https://www.autismspeaks.org/agre |
| 06 | NRGR | https://www.nimhgenetics.org |
| 07 | GEO | https://www.ncbi.nlm.nih.gov/geo |
| 08 | SSC | https://www.sfari.org/resource/simons-simplex-collection |
| 09 | Simons VIP | https://www.sfari.org/funded-project/simons-variation-in-individuals-project-simons-vip |

However, concerning class labels, ASD datasets are imbalanced in measuring performance evaluation (such as sensitivity, specificity, error rate, area under the curve (AUC), and UAR). Imbalanced datasets are mostly utilized by the research works for

the analysis of autism. Since analytical efficiency is reduced in these datasets, hence the need for more reliable and valid datasets becomes essential.

### 2.2.2 Imbalanced Autism Dataset and Procedures to be Addressed

For the privacy and integrity of data, imbalanced data should be processed in order to utilize properly. Researchers attempt to minimize this concern by using various methods such as over-sampling, under-sampling, stratified cross-validation, and integration of datasets from multiple sources. The following are details and descriptions of the methods:

a. Over-sampling: Attempts to reduce the imbalanced class label issue by reproducing the initial minority class instances (non-ASD data) are known as over-sampling. Researchers generally dislike this approach because it is time-consuming, needs high computing resources, and can overfit the training dataset. For example, Bone et al. (Bone et al. 2016) worked on balancing the class labels in the input dataset, where the ratio of ASD to non-ASD instances was 1:3. No instances were lost from the original dataset. Thus, this sampling method is commonly used in medical datasets involving imbalance difficulties.

b. Under-sampling: The technique by which data instances are removed from the majority class (non-ASD) to balance the proportional dissemination of input data related to the class label is known as the under-sampling method. In this sampling method, the intention is to provide a reduced amount of information given by the predictive model, which is a critical issue for prediction and decision making because under-sampling deduces real data instances from the majority class. Over-sampling and under-sampling approaches with a preliminary clustering step seem to be a way forward, with the assumption that data duplication or data removal would be innovative more than absolutely random. Researchers categorize their input dataset into *N* clusters in the under-sampling method. The proportion of majority class instances to the minority class is used to choose instances

that may be utilized for the training stage from various groups based on the computed proportion.

c. Other methods of under-sampling use *k*-nearest neighbors (KNN) from supervised learning. These methods are suitable for decreasing randomization in the sampling process. In these methods, most class instances are taken from various subsets of data depending on a distance function metric.

d. Stratified cross-validation: To address the problem of the imbalanced dataset, Duda et al. (M Duda et al. 2016) proposed stratified cross-validation and under-sampling methods, which enabled their classification model to learn 90% and 10% features from the training and testing datasets, respectively. These methods were applied in a total of 10 different phases. They achieved a proportion of ASD to attention deficient hyperactivity disorder (ADHD) rate of 1.5:1 in every sample set. They randomly employed 10 samples in the under-sampling technique for the majority group of ASD in both training and testing sets to get this outcome.

e. Integration of datasets from numerous resources: Class labels of imbalanced data are minimized, and ASD diagnostic model performance is simplified to integrate data from various resources. It has significance in removing similar features before integrating cases and controls from multiple resources. In addition, codes in the ASD predictive tools are related, such as the modules in ADOS-R. Consequently, an appropriate tactic is to remove significant similarities among features before integrating datasets. The aim is to get dissimilarities in the new integrated features so that these dissimilar features can be measured obviously as the class at the time of feature selection and diagnosis.

f. Finally, regarding the evaluation metrics for constructing an ASD analysis model, divergence occurs when processing imbalanced data. For example, Kosmicki et al. (Kosmicki et al. 2015) presented classifier integrity using

classification accuracy as an evaluation metric. On the contrary, Duda et al. (M Duda et al. 2016) argued that measuring the evaluation from imbalanced datasets is not proper. Instead, UAR is a more appropriate metric that can integrate ASD and non-ASD recall.

However, secured balanced datasets are most necessary for the proper identification of ASD, so that researchers, medical personnel, data analyst and others can utilize those data properly for their own purpose.

### 2.2.3 Different Autism Sample Data with their Various Types are Utilized by Researchers

There are different types of data regarding autism which are utilized by researchers in various kinds of research. These are illustrated broadly below.

Four types of autism data, such as autism-adolescent, autism-child, cryotherapy, and immunotherapy, were employed in a privacy preservation framework for concealing the information where authors utilized artificial bee colony algorithm (Mewada et al. 2020). In another research work of the same author, where he utilized six types of datasets, for example, 198 WPBC, 303 Heart, 704 Autism-Adult, 292 Autism-Child, 90 cryotherapy, and 90 immunotherapies, but in this article, he used the K-Means++ algorithm (Mewada 2021). Through actions like swapping, altering, and deleting, this strategy transforms the original dataset into a protected dataset. To calculate the horizontal partitioning without revealing the information about the cluster centres, the K-Means++ technique was used in this study. To ensure the privacy and semantics of the data, noise is then added to the database. Additionally, the seed function is employed to safeguard the original databases. Utilizing a number of benchmark medical datasets, the efficacy of the suggested method is assessed.

Autism microarray data is also used by many authors. One of them utilized an autism microarray dataset from the well-known public repository gene expression omnibus (GEO), launched by the national centre for biotechnology information (NCBI), which consisted of 146 observations (samples) and 54,613 genes (features) (Hameed et al. 2017). They discussed gene expression from the ASD dataset. The data

from gene expression patterns was used by other authors, where they considered characteristics from spatiotemporal gene expression patterns in the human brain, gene-level constraint metrics, and other gene variation traits for ASD risk genes (Lin et al. 2020).

Structural magnetic resonance imaging (sMRI) data is also used in different studies. There are diverse findings for the structural brain variations in individual having autism. Recent (post-2007) high-resolution (3T) MRI studies studying brain morphology linked with ASD have been included in a comprehensive analysis by Pagnozzi A. M. et al. (Pagnozzi et al. 2018). And the datasets consisted of 2925 complete Social Responsiveness Scale (SRS) score sheets from the AGRE (Geschwind et al. 2001), the SSC (Fischbach & Lord 2010), and the Boston Autism Consortium data repositories were utilized by Duda et al. (M Duda et al. 2017) in their research. They tried to identify autism spectrum disorder (ASD) from the attention deficit hyperactivity disorder (ADHD).

Thabtah illustrated imbalanced datasets in his one of the research (Thabtah 2019). He recommended some fruitful techniques to tackle the imbalanced datasets for proper usage of data in the research. Thabtah looks into current secured and proper experiments in ASD, both clinically and non-clinically which depends on appropriate, suitable and balanced data.

Datasets from various ASD evaluation sheets are unusable for the research. A pre-processing method for separating data object has been introduced by the authors in their research (Pancerz et al. 2015), where they cleaned training data coming from ASD evaluation sheets for utilizing in their research. In the approach, they categorize their procedure into three steps: i) to calculate consistency factor, ii) to divide a set of all training objects (cases) into a subset of unambiguous objects (cases) and a subset of boundary objects (cases). These sheets of ASD are consisted of seventy cases (subject) where every subject was evaluated by questions which is grouped by seventeen area with three hundred attributes. Every attribute had four values (0 → not performed, 25 → performed after physical help, 50 → performed after verbal help/demonstration, 100 → performed unaided).

To apply multidimensional data and big datasets in the research is an important issue. Hyde et al. showed that their method how to allow for multidimensional data, big and rich datasets regarding genetic datasets (Hyde et al. 2019).

A questionnaire dataset related to ASD is used by a researcher (Allison et al. 2012), where they showed how the pre-diagnostic time can be decreased before specialist assessment in medical clinic, because the clinical diagnosis procedure can be time consuming and lengthy for its complex process. They suggest medical staff (clinical staff, care staff, physicians, nurses etc.) to use maximum ten important secured questions for quick referral decision for further ASD diagnostic cases.

Large brain imaging dataset (functional brain imaging data) from the ABIDE (Autism Brain Imaging Data Exchange) multi-site database was utilized by Heinsfeld et al. (Heinsfeld et al. 2018). In this regard, they used 10-fold cross validation as sampling method. Besides, classification must account for extra sources of variation in participants, scanning techniques, and equipment when comparing single-site datasets to multi-site datasets. Such variance introduces noise into brain imaging data, making it difficult to extract signatures from brain activation that may be used to categorise disease states.

Mohamed, S. et al. intended to measure the level of social emotions securely and perfectly by using a instruments like Ages & Stages Questionnaires: Social-Emotional (Malaysia) (ASQ:SE (M)) (Mohamed & Toran 2017). In this purpose, they employed some different ages data, such as, 24 months, 30 months, 36 months, and 48 months autism datasets.

Finally, the different kinds of autism sample data including their various types and techniques which were used to achieve the intended objectives in the numerous research have been summarized in the following Table 2.2.

Table 2.2 Various data utilized in the autism research

| References | Algorithms/ Methods | Objectives | Data Type | Sample Data |
|---|---|---|---|---|
| Mewada et al. (Mewada et al. 2020) | ABC | Data privacy preservation | Screening | ASD-Adolescent, ASD-Child, Cryotherapy, Immunotherapy |
| Bi et al. (Bi et al. 2018) | Random SVM | ASD/TC | rs-fMRI | ASD, TC |
| Duda et al. (Marlena Duda et al. 2016) | ADTree | ASD/non-ASD | ADI-R | ASD, non-ASD |
| Bone et al. (Bone et al. 2016) | SVM | ASD/other DD | ADI-R, SRS | ASD, other DD |
| Heinsfeld et al. (Heinsfeld et al. 2018) | 10-fold cross validation | ASD/TC | rs-fMRI | ASD, TC |
| Lin et al. (Lin et al. 2020) | Logistic regression, SVM, Gaussian kernel, random forest and boosted trees | ASD/ non-ASD | rs-fMRI | ASD, non-ASD |
| Mandala, J. et. al. (Mandala & Rao 2019) | AAP-CSA | Data protection | Screening | ASD-Adolescent, ASD-Child, Cryotherapy, Immunotherapy |
| Duda et al. (M Duda et al. 2017) | Enet, LASSO, SVM, LDA, Ridge regression | ASD/ADHD | SRS | ASD, ADHD |
| Achenie et al. (Achenie et al. 2019) | fNN | ASD/ non-ASD | M-CHAT-R/F | Toddlers |
| Levy et al. (Levy et al. 2017) | Supervised learning | ASD/ non-ASD | ADOS | ASD, non-ASD |
| Mewada (Mewada 2021) | K-Means++ algorithm | Privacy preservation | Screening | ASD-Adult, ASD-Child, Cryotherapy, Immunotherapy |
| Mohamed, S et al. (Mohamed & Toran 2017) | ASQ:SE (M) instrument | Measuring Social-Emotional level | ASQ:SE (M) questionnaire | ASD (24-, 30-, 36- and 48-months) |
| Vaishali et al. (Vaishali & Sasikala 2018) | Binary firefly algorithm | ASD/ non-ASD | ADOS, ADI-R | ASD-Child |
| Wang et al. (C. Wang et al. 2019) | SVM-RFECV | ASD/TD | rs-fMRI | ASD, TD |

## 2.3 AUTISM DATA PRIVACY FRAMEWORKS

This section clarifies privacy first. After that, it elaborates on the numerous privacy techniques or methods of different frameworks in the next section broadly.

### 2.3.1 Data Privacy

Data privacy is a crucial issue nowadays as various types of data are being aggregated enormously every moment. It implies to the dealing with the capability of an individual or an organization that determine what types of data into the transmission line should be distributed with each other. It is considered as an important issue of information sharing.

The very basic concepts of privacy are mentioned in the following Table 2.3.

Table 2.3 Basic concepts about privacy

| **Privacy** |
| --- |
| ▪ Privacy refers to the capacity to control where and how personal information is shared. |
| ▪ The proper usage of user's information is defined as privacy. |
| ▪ Privacy is maintained by the user's consent to share his/her information to others. |
| ▪ It is about the right of patients to keep their personal information private from third parties. |

Therefore, it is obvious that protection of data is inadequate to address without privacy issue.

### 2.3.2 Data Privacy for Autism Dataset in Different Frameworks

In the autism dataset, privacy of data is one of the major issues which is most important concern to the researchers right now. Because a huge number of medical data are being increased day by day in the health data storage in the world. As a result, researchers are contributing to develop their own frameworks or models which maintain privacy issue of data. In purpose to do this work, they have applied various algorithms and techniques for better performances into their frameworks or models.

Eye movements encode powerful information about psychological components of a person, which could be a key to distinguish the traits of ASD. Due to the shortage of private and open datasets, research in this area is still inadequate. In this work, the authors represented an open dataset of eye movements of ASD children, which was consisted of 300 natural scene images and the related eye movement data accumulated from 14 healthy children and 14 children with ASD in a secure way (Duan et al. 2019). Depending on this dataset, researchers could examine the visual behaviours of children with ASD and construct particular model to encourage research in the related research area and to identify the individuals with ASD. So, this type of data should be maintained the privacy.

Other authors developed a cost-effective privacy framework that suits e-health authentication and data transmission (Boonyarattaphan et al. 2009). They proposed two risk adaptive authentication methods that are appropriate for various circumstances of e-health services. The benefits of authentication methods are high service quality data protection. And finally, they demonstrated the applicability of different encryption algorithms such as DES (Data Encryption Standard), 3DES or TripleDES, AES (Advanced Encryption Standard), Blowfish, IDEA (International Data Encryption Algorithm) and RC4 (Rivest Cipher 4) with optimum key length that provides a cost-effective result for obtaining e-health services.

Sivan, R. et al. (Sivan & Zukarnain 2021) reviewed various types of data privacy models for e-health database. They highlighted the weakness and strengths of those models. Firstly, they highlighted some important e-health protection issues, for example, confidentiality, integrity, availability, data violations, wrong fix, lack of privacy technologies, account hijacking, insider threat, etc. After that they discussed some solutions in e-health systems: cryptographic solutions (such as public key encryption (PKE), symmetric key encryption (SKE), identity-based encryption (IBE), searchable encryption (SE), and attribute-based encryption (ABE), proxy encryption, homomorphic encryption, broadcast encryption programs, qualified encryption, blockchain-based encryption, searchable symmetric encryption (SSE), and access control manager (ACM) methods to protect cloud-based data in the e-health architecture.

Mandala, J. et. al. developed a model for privacy of autism data (Mandala & Rao 2019). They suggested on introducing an efficient sanitizing technique for hiding the sensitive data. For hiding the confidential medical data, an optimal key was generated by Adaptive Awareness Probability-based Crow Search Algorithm (AAP-CSA). In their study, authors applied the crow search algorithm, a unique metaheuristic optimization algorithm, into their own model. This algorithm is built on modelling the intelligent performance of crow flocks.

An e-health framework (eHF) was proposed by another author (Ding & Klein 2010). It is a complete development platform for electronic health care applications. It provides a toolkit for designing, developing, building, deploying, and maintaining eHF-based applications in a development environment. The model is also known as model-driven application-level encryption technique where they combined the flexibility of application-level encryption and the transparency of database-level encryption. The safety module was included with authentication, authorization, auditing, user management and application-level encryption. They claimed that their framework was a more comprehensive and protective platform for the development of electronic health care solutions.

Another model for data privacy is Global Patient Identification Technique (GPIT), where researcher anonymized recognizable private data of patients while maintaining record linkage in integrated health repositories to accelerate knowledge discovery development (Khan & Hoque 2016). They have applied encrypted mobile number, gender and name-value of patients for generating Global Patient Identification Key.

In the article (Majeed 2019), researcher recommended a new anonymization scheme of data privacy for e-health records. This new approach built on fixed intervals for generalizing the numerical attributes of e-health records and is solely based on data values. Therefore, he claimed that this approach will provide necessary knowledge to research, data publishers and key-players in the privacy area for modelling and creating more secured solutions for protecting the privacy of one's publishing data. He also claimed that it prevents from identity disclosure even faced with adversaries having

pertinent background knowledge and improves data privacy and utility in privacy preserving data publishing.

In another research work, a researcher developed personal health data management system as a privacy and security model (Wu et al. 2016). They applied this privacy protection scheme on the W-iPCN, android smartphone and the central medical server to test their ability to process and transmit raw sensor data in real-time. This aim of this scheme is to observe the performance of data encryption and decryption, and real-time feasibility to remotely monitor patients' health.

Autism Management Platform (Linstead et al. 2016) was a client-server architecture that included a mobile app along with an analytics engine, a reliable database, and a web client where caregivers and clinicians captured multimedia autism data, then they disseminated, navigated, and analysed these. After that the system delivered pertinent responses within short times, by filtering and learning data in purpose of reducing redundancy, to the guardians of individuals with children.

In other two articles of Khaloufi, H. et al. (Abouelmehdi et al. 2018; Khaloufi et al. 2018), the authors mentioned the privacy issue for big healthcare data. They, firstly, discussed various threats and attacks like spoofing, spamming, content-based attacks, data mining-based attacks, re-identification threats, phishing etc. at various phases like, data collection phase, data transformation phase, data modelling phase, knowledge creation phase and suggested important ways to maintain privacy and security of healthcare data finally.

Bernstein, D. J. (Bernstein 2005) discussed about brute-force attack. He described various attacking process and numerous privacy and safety issues. Mainly attackers build different length of keys to the attacks like the standard key-search circuit, the variant key-search circuit, the standard key-search machine, the variant key-search machine, and fancier serial attacks etc. and suggest forming input space separation and larger key which have much larger benefits and much smaller costs.

Researchers applied different types of sanitizing techniques to conceal the sensitive information. Data sanitization (Edgar 2004) is a technique for masking sensitive data in development databases which can be accomplished by replacing it with fake data of a comparable type that seems to be real. There are different kinds of sanitizing techniques, for instance, encryption/decryption, NULL'ing Out, number variance, shuffling records, substitution, masking data, etc. and are utilized in various framework and models.

Thomas, R. M. et. al. worked to tackle the missing data, small size of dataset and heterogeneity (Thomas et al. 2020). In this case, they applied imputation methods for missing data because imputation aims to fill in the values of missing data while preserving the characteristics of their distribution and maintaining the relationships to other variables. It gives enormous advantages in the context of clinical experiments, for instance, by lessening the bias combined with the insertion of patient with complete data and expanding the use of the available information. And for small size of dataset and heterogeneity, they used data augmentation, transfer learning, simulation-based augmentation, data efficient learning techniques.

High-dimensional datasets with redundant, nosy and irrelevant features decrease classification performance and increase the computational cost. So, lowering the data dimensionality and choosing only the very appropriate features is done by optimization algorithms. Grey Wolf Optimization (GWO) (Abdel-Basset et al. 2020; Al-Tashi et al. 2020) is one of them which is taking these challenges in particular for datasets with a huge quantity of features. Here, Al-Tashi, Q. et al. (Al-Tashi et al. 2020) recommended in their work to investigate the usage of angle modulated function or additional binary operators with GWO to resolve this problem, whereas Abdel-Basset, M. et al. (Abdel-Basset et al. 2020) applied GWO integrated with a two-phase mutation to solve this issue based on the wrapper methods. Since GWO have a numerous usability so in the study (Mirjalili et al. 2020), Mirjalili, S. et al. discussed various GWO like discrete GWO, constrained GWO, multi-objective GWO, hybrid GWO and examines the application of the GWO variants in obtaining the optimal model for a ship propeller.

In a work of Payne, K. L. et al. (Payne et al. 2019), they demonstrated the link between cybercrime and autistic-like characteristics, or autism. They discovered that higher degrees of sophisticated digital abilities were responsible for about 40% of the link between autistic-like characteristics as well as cyber-dependent crime. In this regard, much protection should be taken for increased data sharing through the internet which can decrease scattering malware, hacking medical data, stealing personal information, damaging reputation distributed denial of service (DDoS) attacks committed by individual with autistic traits.

### 2.3.3 Summary of the Previous Research Works

For the purpose of privacy for autism sensitive data, different models have applied different methodologies to achieve their desired objectives. The following Table 2.4 shows some state-of-the-art frameworks or models for privacy and protection of health sensitive data. In this regard, I have highlighted specific datasets, identification of problems, along with the various approaches used by the frameworks, contributions, and strengths.

Table 2.4 Comparison among different data privacy models

| References | Datasets | Problem Identification | Approach/Methodology | Contribution/ Strength |
|---|---|---|---|---|
| Abouelmehdi K. et al. (Abouelmehdi et al. 2018) | ▪ Big healthcare data. | ▪ Threats and attacks in various phases of big healthcare data lifecycle. | ▪ Authentication, encryption, data masking, access control, monitoring and auditing.<br>▪ k-anonymity, l-diversity, and t-closeness to enhance de-identification.<br>▪ HybrEX (such as map hybrid, vertical partitioning, horizontal partitioning, hybrid) | ▪ Several risks and assaults at every stage of the big data life cycle regarding healthcare.<br>▪ Suggesting defences and strategies available accordingly. |
| Khaloufi, H. et al. (Khaloufi et al. 2018) | ▪ Medical data. | ▪ Privacy issues in different phases for healthcare data. | ▪ Administrative rules; filtering and classifying; clustering, classification, association several ensembles learning techniques; defence compliance, verification. | ▪ Critical privacy and protection treats identified in the life cycle of big data in regard with healthcare sector.<br>▪ Feasible techniques and solutions, therefore, against those numerous threats and attacks of every phases accordingly. |
| Khan, S. I. et al. (Khan & Hoque 2016) | ▪ Patients' health data. | ▪ Insecure national health data warehouse. | ▪ Encrypted mobile number, gender and NAMEVALUE of patients. | ▪ Global Patient Identification Technique (GPIT).<br>▪ Can anonymize identifiable private data of the patients. |
| Linstead E. et al. (Linstead et al. 2016) | ▪ Autism health dataset | ▪ Data redundancy of autistics patients.<br>▪ Not rapidly feedback to the patients' necessary queries and medication. | ▪ Aggregating, filtering, learning and mining patients' data to deliver feedback to the guardians with autism.<br>▪ HTTP/SSL protocol through industry-standard 128-bit encryption.<br>▪ A 1024-bit token along with the session which is cryptographically protected. | ▪ A large scale of patients' data.<br>▪ Quick responses to the patients' queries. |

to be continued …

… continuation

| | | | | |
|---|---|---|---|---|
| Majeed, A. (Majeed 2019) | ▪ Adults' health data. | ▪ E-health privacy problem | ▪ For generalizing numerical attributes, specified intervals are used, and original values are replaced by averages. | ▪ Attribute-centric anonymization scheme.<br>▪ Disclosure of identity is protected by the scheme even when faced with adversaries having pertinent background knowledge.<br>▪ Recommended approach improves data privacy and be used to publish data while maintaining privacy. |
| Mandala, J. et al. (Mandala & Rao 2019) | ▪ Autism Dataset (Adolescent, Child, Cryotherapy, and Immunotherapy dataset) | ▪ Privacy of autism data for the healthcare data. | ▪ Sanitizing approach.<br>▪ Crow Search Algorithm (CSA). | ▪ Working with meta-heuristic algorithm, such as Adaptive Awareness Probability-based CSA (AAP-CSA).<br>▪ Performance of this approach is superior to PSO, GA, DE, CSA designs in regard of objective functions, $c_1$, $c_2$ and $c_3$. |
| Sivan R. et al. (Sivan & Zukarnain 2021) | ▪ Healthcare data in the cloud | ▪ Privacy threats in e-health data | ▪ Reviewed Public Key Encryption (PKE), Identity-Based Encryption (IBE), Searchable Encryption (SE), and Attribute-Based Encryption (ABE). | ▪ Proposed a number of methods to protect cloud-based data in the e-health architecture. |
| Wu, P. et al. (Wu et al. 2016) | ▪ ECG Sensor Data. | ▪ Vulnerable Medical Dataset. | ▪ W-iPCN (Wireless Intelligent Personal Communication Node), Android smartphone and the central medical server. | ▪ Personal Health Data Management System (PHDMS).<br>▪ A secured data transmission mechanism in real-time. |

## 2.4 TRADITIONAL META-HEURISTIC ALGORITHMS

Recently, meta-heuristic algorithms are becoming popular day by day. A meta-heuristic algorithm is an algorithmic framework that gives a set of principles or developing methods to solve the problem at higher level. There are numerous kinds of meta-heuristics algorithms which are being used in solving the real-life problems. There are various meta-heuristics algorithms, such as Artificial Bee Colony (ABC), Crow Search Algorithm (CSA), Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Glowworm Swarm Optimization (GSO), Jaya Algorithm (JA), Shark Smell Optimization (SSO), Cuckoo Search Algorithm (CSA), Gravitational Search Algorithm (GSA), Differential Evolution (DE), Grey Wolf Optimization (GWO). Though these algorithms are employed in the area of engineering, business, marketing, and etc. But most popularly, these algorithms are utilized vastly in the healthcare sector for privacy of medical data. This research has employed two algorithms PSO and GWO as combinedly. These algorithms are illustrated as below:

### 2.4.1 Particle Swarm Optimization (PSO) Algorithm

A meta-heuristic, particle swarm optimization (D. Wang et al. 2018) is a stochastic optimization method which depends on swarm. This optimization is recommended first by Kennedy and Eberhart in 1995 (Kennedy & Eberhart 1995). PSO algorithm is simulated by collective animals' behaviour by the way of grouped insects, herds, schools of fish or flocks of birds, etc. It is a swarm-based search method in where every individual is referred to as a particle and is described as a potential solution to the optimum issue in D-dimensional search space, and it can remember the swarm's and its own ideal positions, along with their velocity. The information of the particle in every generation is merged to adapt the velocity of every dimension, which is then utilized to determine the particle's new location. Particles in the multi-dimensional search space continually alter their states until they achieve optimal or the best state, or they go beyond the calculation limitations. The objective functions establish distinctive correlation amongst separate dimensions of the problem space. A flowchart of this optimization is depicted in Figure 2.1.

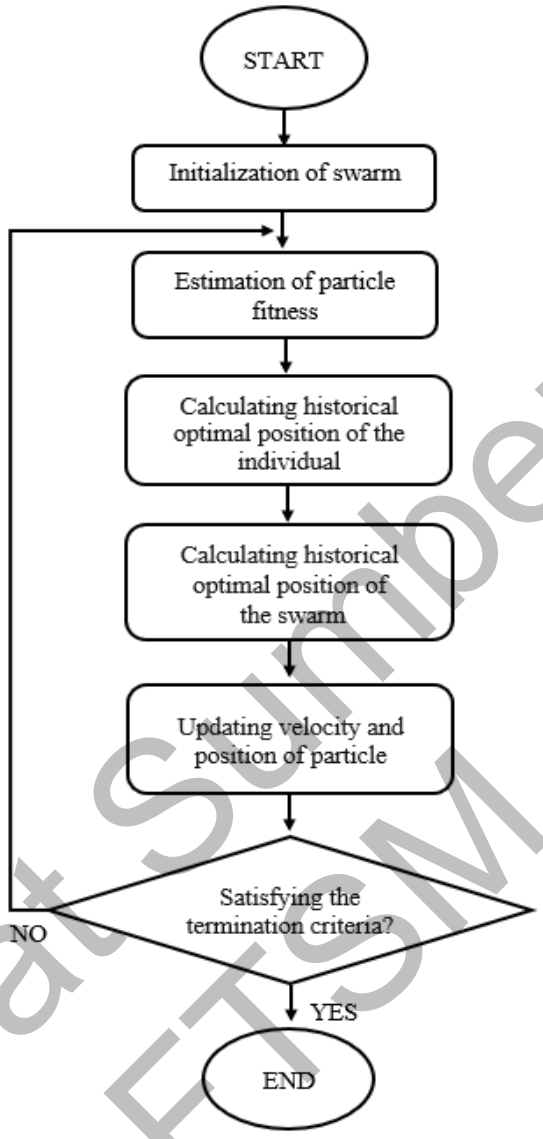



Figure 2.1 Flowchart of the particle swarm optimization algorithm adapted from (W. Li et al. 2021)



This method has been shown to be a successful optimization technique in a number of studies. It has significant advantages, such as

- PSO has lowered number of parameters containing two acceleration coefficients and merely inertia weight factor compared to another competing heuristic optimization techniques.
- It has straightforward coding implementation, computational effectiveness, easy concept, and robustness to regulate parameters.

- This method poses stable convergence features and be able to produce high-quality solutions within quicker calculation time over another stochastic approaches.
- It is a derivative-free method and less sensitivity to the environment of the objective function in comparison with another heuristic processes and other conventional mathematical methods.
- It appears to be slightly less dependent of a set of initial points in comparison with another evolutionary techniques, suggesting that convergence algorithm is robust.

### 2.4.2 Grey Wolf Optimization (GWO) Algorithm

Another meta-heuristic algorithm, grey wolf optimization (Mirjalili et al. 2014) is a new optimization algorithm which is first proposed by Mirjalili et al. It is based on the natural hierarchy of grey wolves' leadership and hunting mechanism. For modelling the leadership structure, four sorts of grey wolves are used, for instances, alpha, beta, delta, and omega. The three major phases of hunting are carried out, such as seeking for prey, surrounding prey, and attacking prey. A hierarchy of grey wolves are given in Figure 2.2 where dominance declines from top to bottom.

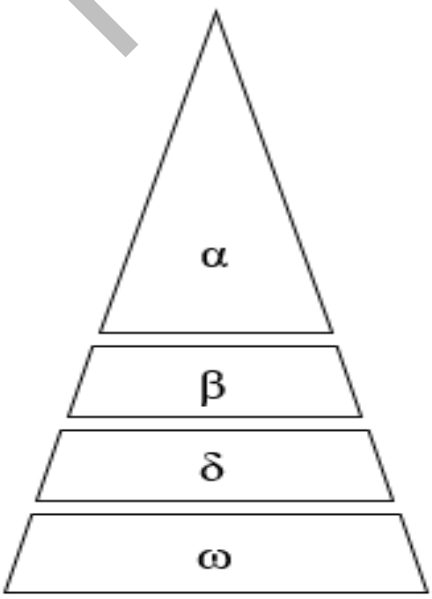



Figure 2.2 Hierarchy of grey wolves adapted from (Y. Li et al. 2021)

There are some advantages in this optimization algorithm as below,

- The limited number of search parameters and user-selected parameters are valuable benefits of GWO algorithms which is reflected in a variety of applications, like solutions to standards mostly used to assess optimization algorithms.
- This technique demonstrates its capability in resolving numerous optimization challenges, for example, feature selection methods for classification, obtaining the optimal model for a ship propeller with a lowered user knowledge and fair comparison with related metaheuristics, and etc.

A flowchart of grey wolf optimization is shown in Figure 2.3.

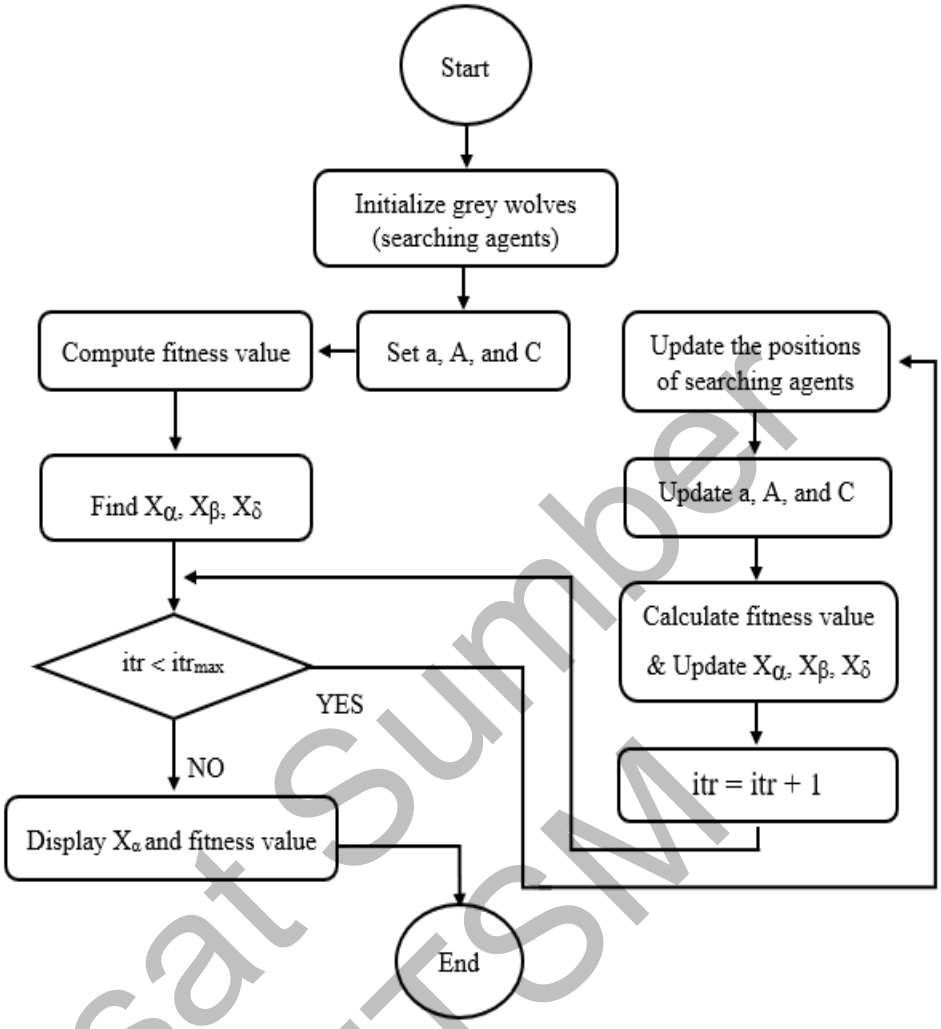



Figure 2.3 Flowchart of the grey wolf optimization algorithm adapted from (Y. Li et al. 2021)



For the usage of this optimization algorithm, it is illustrated in the proposed framework in Chapter IV and Chapter V.

## 2.5 DISCUSSION ON VERY RELEVENT EXISTING WORKS REGARDING DATA SANITIZATION AND DATA RESTORATION

This section examined and analyzed the very relevant works in order to find out research gap and concerns that need to be addressed in respect of data sanitization and data restoration for sensitive medical data privacy. In order to sanitize the sensitive data through sanitization process, the relevent works have been discussed in section 2.5.1. On the other hand, the relevent works regarding data restoration have been illustrated in section 2.5.2.

### 2.5.1 Reseach Gaps regarding Data Sanitization

The study, in this section, analysed the related studies concerning data sanitization for the privacy of sensitive data and their characteristics, methods or techniques, and challenges that required to be focused.

An author developed a privacy model for hiding the sensitive information of medical data with the help of an artificial bee colony-based (ABC-based) model (Mewada et al. 2020). The ABC-based model creates an optimal key for anonymizing sensitive information, and the same key is used to restore information. They also considered four threats, for example, known cipher attack (KCA), known-plaintext attack (KPA), chosen cipher attack (CCA), and chosen-plaintext attack (CPA), for the validity of the performance of their suggested approach. However, this study assessed the fitness functions, such as information hiding failure rate, information preservation rate, and degree of modification rate. They showed the information preservation rate to be minimized for their work that should not be suitable objective. So, this problem should be address.

Depending on adaptive awareness probability with a meta-heuristics algorithm, namely the crow search algorithm, a researcher improved a data preservation method for medical data (Mandala & Rao 2019). The suggested framework deals with the method of data sanitization to mask sensitive data. In comparison to other existing techniques, the effectiveness of the suggested system was observed, and it was found that their suggested system offers rigorous and efficient results for the privacy of autism data. In this case, the objective functions that refer to performance matrices and are utilized in this article need to address also.

Data privacy is also significant concerns for the cloud computing environment because this environment is full of access to various data, files, and applications. For its advantages, the cloud is exploited in the healthcare sector vastly. For example, in a work (Alphonsa & Amudhavalli 2018), researcher developed a secure model named GMGW to sanitize sensitive information of heart disease data based on the cloud system. And in the same way, authors in numerous studies, (Ahamad et al. 2022; Balashunmugaraja & Ganeshbabu 2020; Lekshmy & Rahiman 2019), developed privacy models

individually applying different algorithms on the cloud computing system for data privacy and compared the performances of their models with the conventional algorithms for more improvement.

Abidi MH et al. established a secured data transmission model as Whale with New Crosspoint-based Update (WNU), which is the update version of the Whale Optimization Algorithm (WOA), in supply chain management along with blockchain technology (Abidi et al. 2021). They also evaluated their model concerning four research issues like false rule (FR) generation, information preservation (IP) rate, hiding failure (HF) rate, and degree of modification (DM).

Ochôa, IS et al. also apply blockchain technology to protect users' personal data by using three blockchains to confirm privacy, security, and trust in their architecture (Ochôa et al. 2020). They utilized sidechains for scalability and adaptability of their system, whereas another researcher applied an optimal key in their proposed model for the privacy-preserving data mining (PPDM) technique using an opposition intensity-based cuckoo search algorithm (Shailaja & Rao 2020). They also assessed their model by FR, IP, HF, and DM.

In also the various works, (Han et al. 2020; Renuga & Jagatheeshwari 2018; Revathi et al. 2018; Sahi et al. 2016), authors built privacy models independently applying various algorithms on the cloud computing system. They also compared the performances of their models with the conventional approaches for enhancement.

Liu Y et al. introduced a new reversible data hiding strategy based on the region of interest (ROI) in encrypted medical healthcare images of the patients (Y. Liu et al. 2016). A data owner primarily divides an original diagnostic image into the region of interest (ROI) and the region of non-interest (RONI). The encryption key was subsequently used in anonymizing the images in this regard. The least significant bits (LSB) of the encrypted ROI and electronic patient record (EPR) were concatenated by a data hider. After that, this concatenated data was embedded into the encrypted image by the LSB substitution technique. By the data-hiding key, the receiver retrieved the embedded data contained in the encrypted medical image. If the recipient holds the

encryption key, directly decrypting the encrypted medical image, a medical image similar to the original image could be produced. But suppose the recipient had both keys (data-hiding key and encryption key), embedded data could be retrieved without any mistake, subsequently extracting that embedded data ROI could be retrieved without any flaw.

In a research article (Zhang et al. 2018), authors established a Privacy-Aware Smart Health (PASH) access control system. The key ingredient of their system was a large universe ciphertext-policy attribute-based encryption (CP-ABE), whose access strategy was somewhat secret. The access strategy in the encrypted s-health records (SHRs) was that the attribute values were hidden, but only the attribute names were exposed. Indeed, attribute values hold much more sensitive knowledge than generic attribute names. In specific, PASH conducted an effective SHR decryption test that involves a limited number of bilinear pairings. Moreover, the attributes universe could be infinitely large, and public parameters were small and constant in size. From the analysis, they claimed that PASH was a completely secured one in standard frameworks.

Sharma U et al. recommended two parallelized methods called PGVIR and PHCR (U. Sharma et al. 2020). These approaches were applied to the spark framework, which manipulates the data so that no sensitive data could be retrieved at the time of ensuring the utility of sanitized data. Taking the standard dataset through the experiment, they manifested that PGVIR was better scalable while PHCR ensured dataset quality. Sharma S et al. (S. Sharma & Toshniwal 2020) suggested an approach that optimally reduced the side effect of the hiding process on non-sensitive data, provided a balance between knowledge and privacy, and successfully regulated the rapid increase of data volume.

Though some research works (Alesawy & Muniyandi 2016; Idowu & Muniyandi 2019; Mottalib et al. 2016; Rahman et al. 2019, 2021; Usman et al. 2021) utilized cutting-edge technologies for diagnosing different human disabilities and illnesses, maintaining robustness and accuracy, there was also an imperative urge to data privacy.

It is noted from the comprehensive literature survey that a vast number of algorithms with advanced techniques were generated for anonymization (Bostanoğlu & Öztürk 2020; Iwendi et al. 2020; X. Liu et al. 2020; Zaman et al. 2017). It is possible to describe these algorithms as single objective, multi-objective and restricted algorithms. These algorithms aim to retain information or data that is sensitive.

But none of the algorithms will ensure the protection of knowledge as required for usefulness and privacy. There is, however, a need for an efficient model of anonymization to protect medical records. The latest trends have demonstrated that confidential knowledge or sensitive information is being maintained often by meta-heuristic algorithms. The purpose of these algorithms is to produce an optimal key for the method of sanitization. These algorithms are shown to have better outcomes in comparison with conventional algorithms. Some of the research often use the k-anonymity and query to fix the privacy issues. But these techniques take a great deal of time as well as resources for computation.

Therefore, in this study, an attempt is made to establish an optimal key for protecting privacy using PSO and GWO algorithms for the sanitization process. The following Table 2.5 represents a short summary of different models or methods that utilized cutting-edge techniques or algorithms, especially in data sanitization along with their main features and challenges.

Table 2.5 Significant features and challenges of several privacy frameworks for sanitizing data.

| Authors/ References | Techniques/ Methods | Attributes/Characteristics/ Features | Challenges |
|---|---|---|---|
| Abidi et al. (Abidi et al. 2021) | ▪ WNU algorithm | ▪ Uses the features of supply chain networks depending on blockchain technology.<br>▪ Evaluated using HF rate, IP rate, FR generation, and DM. | ▪ The selection of optimal key in the key extraction phase is the most significant challenge. |
| Ahamad et al. (Ahamad et al. 2022) | ▪ J-SSO algorithm | ▪ Usage of the beneficial features of JA and SSO algorithm collectively.<br>▪ Work with various datasets, like air quality, concrete data, heart disease, superconductivity, and wholesale customer data.<br>▪ Efficient cloud data privacy preservation model with the data sanitization and restoration approach.<br>▪ Work with multi-objective functions involving various parameters like degree of modification, hiding ratio, and information preservation ratio. | ▪ The inaccurate and inefficient offering of privacy measures for data transmissions and operations in the cloud.<br>▪ Susceptible data by untrustworthy cloud environment providers. |
| Alphonsa et al. (Alphonsa & Amudhavalli 2018) | ▪ GMGW algorithm | ▪ Preserve sensitive healthcare data.<br>▪ Hybridization of GA along with the GSO algorithm.<br>▪ Analyze the effectiveness of sanitization, restoration, analysis on convergence, and key sensitivity statistically.<br>▪ Running parallel computation is simpler.<br>▪ Possess higher probability and proficiency in achieving the global optima. | ▪ Vulnerable unencrypted data gathered at the remote cloud storage server.<br>▪ Internal and external threats launched by unreliable cloud service providers and suppliers.<br>▪ It can converge prematurely and be trapped into a local minimum, especially with complex problems. |
| Balashunmugaraja et al. (Balashunmugaraja & Ganeshbabu 2020) | ▪ CI-LA algorithm | ▪ Perform multi-objective functions, including different parameters.<br>▪ Analyze the effectiveness of sanitization, restoration, analysis on convergence, and key sensitivity statistically. | ▪ Hard to configure the keys accurately.<br>▪ Network connection dependency.<br>▪ Essential to keep updating the new software. |

to be continued …

… continuation

| | | | |
|---|---|---|---|
| Han et al. (Han et al. 2020) | ▪ CloudDLP | ▪ Browser-based applications on cloud storage. | ▪ The outside enterprise in cloud services can easily unveil documents or sensitive data in images. |
| Lekshmy et al. (Lekshmy & Rahiman 2019) | ▪ ABC algorithm<br>▪ | ▪ Available users are clustered in distributed computing.<br>▪ Among the users from each group, a user known as a helper user transmits data nominated via the service provider.<br>▪ Evaluated in terms of a few factors (such as clustering accuracy, processing time, and data transmission time). | ▪ Big data sets are not encrypted in a distributed system by using the kernel k-means algorithm for encryption. |
| Renuga et al. (Renuga & Jagatheeshwari 2018) | ▪ GSA algorithm | ▪ Lower execution time, hiding failure, maximum dissimilarity value in comparison with the existing technique. | ▪ Possibility of malicious threats in the sensitive information gathered in the cloud. |
| Revathi et al. (Revathi et al. 2018) | ▪ BS-WOA algorithm | ▪ Involve a small number of parameters and lack of local optima entrapment for resolving clustering problems. | ▪ Hard to keep up with the privacy of every database. |
| Shailaja et al. (Shailaja & Rao 2020) | ▪ OI-CSA algorithm | ▪ Gives superior runtime and scalability. | ▪ Essential to increase privacy-preserving data mining. |

### 2.5.2 Reseach Gaps regarding Data Restoration

In this section, the study analysed the related studies concerning data restoration for the privacy of sensitive data and their characteristics, methods or techniques, and challenges that necessary to be focused.

For a distributed clustering, sanitized data are transferred to the cloud service provider by a helper user and performance is assessed by means of transmission time, processing time and clustering accuracy (Lekshmy & Rahiman 2019).

To address data privacy, Alesawy O et al. (Alesawy & Muniyandi 2016) implemented a method using Elliptic Curve Diffie-Hellman (ECDH) keys with Artificial Neural Network (ANN) and Genetic Algorithm (GA). They decreased encryption and decryption time, which were the challenges behind the ECDH.

For privacy-preserving utility mining (PPUM), a sanitization method called Improved Maximum Sensitive Itemsets Conflict First (IMSICF) algorithm (X. Liu et al. 2020) was suggested, in which maximum conflicts in victim itemsets from sensitive itemsets are tallied to be concealed. To minimize the side effects on non-sensitive information, this approach picks transactions with a small number of non-sensitive itemsets and a high utility of concealed sensitive itemsets for modification.

In the work (Sowjanya et al. 2021), Sowjanya K et al. proposed an improved protocol known as lightweight Elliptic Curve Cryptography based Anonymous Authentication (AA) protocol for the Internet of Medical Things (IoMT). They improved the privacy concerns about the weakness of He et al.'s new AA scheme (He et al. 2017).

At the end, Table 2.6 shows a brief summary of various models or methods that utilized advanced techniques or algorithms, especially in data restoration along with their major characteristics and challenges.

Table 2.6 Methods and processes with various characteristics and challenges of different models for restoration of data

| **Authors/ References** | **Methods/ Algorithms** | **Characteristics/ Attributes** | **Challenges/ Things should be addressed** |
|---|---|---|---|
| Alphonsa et al. (Alphonsa & Amudhavalli 2018) | ▪ GMGW algorithm | ▪ Sensitive healthcare data privacy.<br>▪ Hybridization of two algorithms, GA and GSO.<br>▪ Evaluating performance according to key sensitivity, effectiveness of sanitization, restoration, analysis on convergence, and statistical analysis.<br>▪ Simpler parallel computation,<br>▪ Have better probability and proficiency for global optima. | ▪ Available non-secured information gathered at the cloud server.<br>▪ Numerous threats introduced by unreliable cloud service providers. |
| Lekshmy et al. (Lekshmy & Rahiman 2019) | ▪ ABC algorithm. | ▪ Users are grouped by k-means clustering algorithm.<br>▪ A helper user from each group performs the data transmitting task.<br>▪ Analysing performance by processing time, clustering accuracy, and data transmission time. | ▪ Only kernel k-means algorithm is not able to encrypt big datasets in a distributed system. |

to be continued …

… continuation

| | | | |
|---|---|---|---|
| Liu et al. (X. Liu et al. 2020) | ▪ IMSICF algorithm | ▪ Consider maximal utility count in sensitive itemsets and minimal number of non-sensitive itemsets. | ▪ Other types of sensitive information, for instances, frequent and utility itemset preservation. |
| Mandala et al. (Mandala & Rao 2019) | ▪ AAP-CSA algorithm. | ▪ Hiding sensitive medical data.<br>▪ Using optimal key. | ▪ How long the key will be updated.<br>▪ Number of key ranges. |
| Mewada et al. (Mewada et al. 2020) | ▪ ABC algorithm. | ▪ Anonymising sensitive data.<br>▪ Forming optimal key. | ▪ The objective function, such as information preservation rate needs to address.<br>▪ Not considered for time and space dimensions regarding privacy-preserving. |

## 2.6 SUMMARY

This thesis has taken into consideration the most related works for data privacy. Before reviewing the literatures, a clear discussion on autism dataset, such as sources of autism data, problematic imbalanced data and the techniques to tackle this problem, sample data types, differences between data privacy and data privacy, and so on have been illustrated. After that by reviewing these most related previous works, this study points out the significant findings, such as the state-of-the-art technologies used by the various frameworks, their strengths and contributions, the research gaps and challenges for privacy regarding autism dataset and concludes the significance of privacy issues more importantly. Noted that the subsequent chapter will discuss to fulfil the objectives of this research emphasizing privacy issue into the proposed techniques of the proposed framework.

# CHAPTER III

# METHODOLOGY

## 3.1 INTRODUCTION

This chapter explains and emphasizes the methodological explanation of this research work. But before that it describes the entire working progress phases including data pre-processing phase initially to evaluation. The methodology for this research work illustrates the three main aspects logically, autism datasets with different ages along with the sources. After that this chapter also mentions the implement environments (For example, WEKA Software and Python Programming Languages) by which the recommended framework is evaluated. The recommended framework, namely Enhanced Combined PSO-GWO framework, employs two metaheuristic algorithms, such as particle swarm optimization and grey wolf optimization algorithms, which has been shown in Figure 3.3 along with the sanitization key and main two processes, such as sanitization process and restoration process. Noted that generating optimal key and these two processes are discussed in Chapter IV and Chapter V in details respectively to achieve the objectives of this research.



## 3.2 RESEARCH PHASES

The phases of this research are mainly comprised of three phases for achieving the objectives mentioned in section 1.5. They are data pre-processing phase, data privacy phase, and evaluation phase which are shown in Figure 3.1 as below:

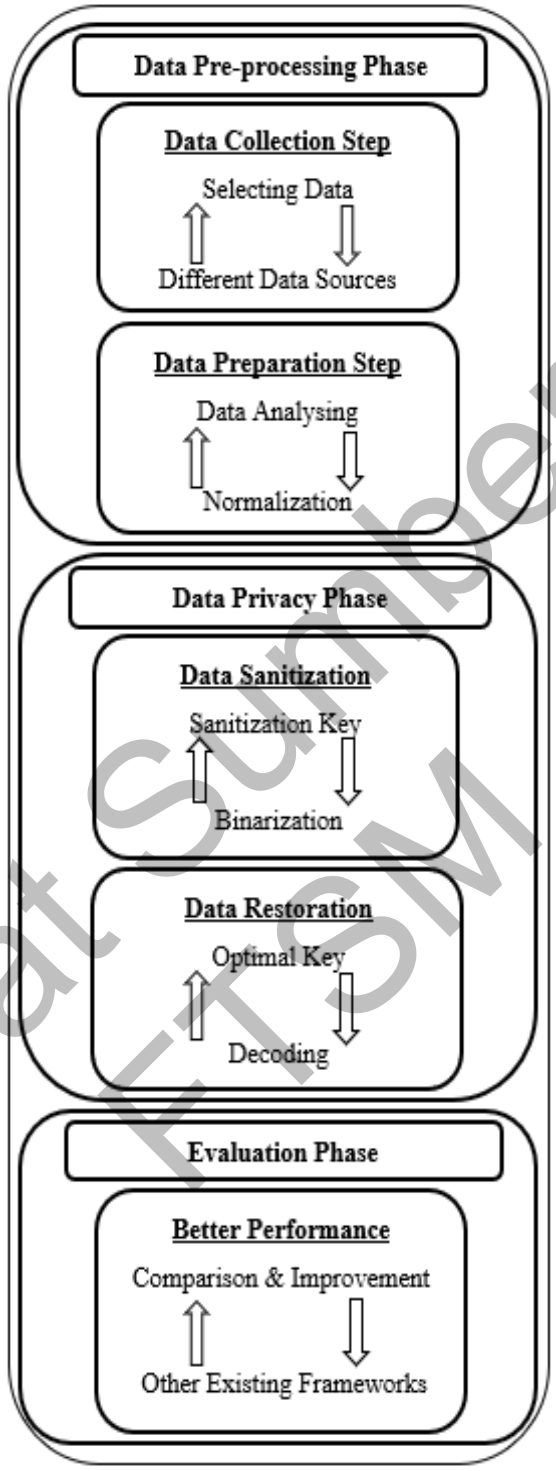



Figure 3.1 Research phases

### 3.2.1 Data Pre-processing Phase

Data pre-processing is an important phase for the research. This phase consists of two steps: data collecting step and data preparation step. For data collecting step, this

research needs to collect datasets for pre-processing. In this regard, autism datasets had been collected from the Centre of Community Well-being and Education, Faculty of Education, Universiti Kebangsaan Malaysia. Autism datasets also can be downloaded from most familiar repositories like Kaggle, ABIDE (Autism Brain Imaging Data Exchange), AGRE (Autism Genetic Resource Exchange), NDAR (National Database for Autism Research), NRGR (NIMH Repository and Genomics Resource), UCI (UC Irvine Machine Learning Repository). However, after collecting autism datasets, it is necessary to do analysis and normalization, discretization, numeric transformation, formatting sparse data, setting class attribute, handling missing data, or duplicate data, etc., at the preparation step for using the datasets. In this phase, support vector machine (SVM) by applying WEKA software is applied for the pre-processing of the autism data.

### 3.2.2 Data Privacy Phase

For the data privacy, the usage of methods and techniques depends on individual researcher. Different researchers apply different methodologies to achieve their goals. Mainly, they apply state-of-the-art methods or technologies for the robustness, and better performance. However, autism sensitive data preservation for this research has been done by generating optimal key and two procedures, for instance, the data sanitization, and the data restoration procedures. To generate the optimal key, the key generation process, key encoding, key transformation, and key extraction, etc. along with the sanitization procedure have been discussed broadly in Chapter IV. On the other hand, restoring the data by restoration procedure discussed first, and also the more effectiveness of the cost function based on varying acceleration constants in the restoration procedure is illustrated in Chapter V.

### 3.2.3 Evaluation Phase

In this phase, the developed framework compares the achieving results with the other models' functionality to achieve of this research objectives successfully. Some objective functions as performance matrices are measured to indicate how well the framework maintain the privacy. This activity involves comparing the objectives of the recommended framework to the results from other existing privacy algorithms or

models and to enhance the recommended framework where necessary for required implementation, that is, this activity can decide whether to iterate back to the framework for further development. Conceptually, such evaluation work might be involved appropriate logical corroboration or evidence. For every objective of this research work, the evaluation phase is illustrated broadly in the individual section of the respective chapter, that is the simulation and analysis, and results and discussions sections comparing to other existing algorithms.

## 3.3 RESEARCH METHODOLOGY

Research methodology refers to a procedure or technique that identifies the problems of a specific topic and process and analyses information to achieve desired goals. There are three primary methodology types, such as qualitative, quantitative, and mixed techniques methods are used by researchers. Research methodology consists of the following key points: defining solutions, designing and development, and implementation and evaluation.

## 3.4 RESEARCH METHODOLOGY FOR THIS RESEARCH WORK

This section explained the research methodology for this research work. The methodology has been illustrated in Figure 3.2 which is modified to fit the demands of this study.

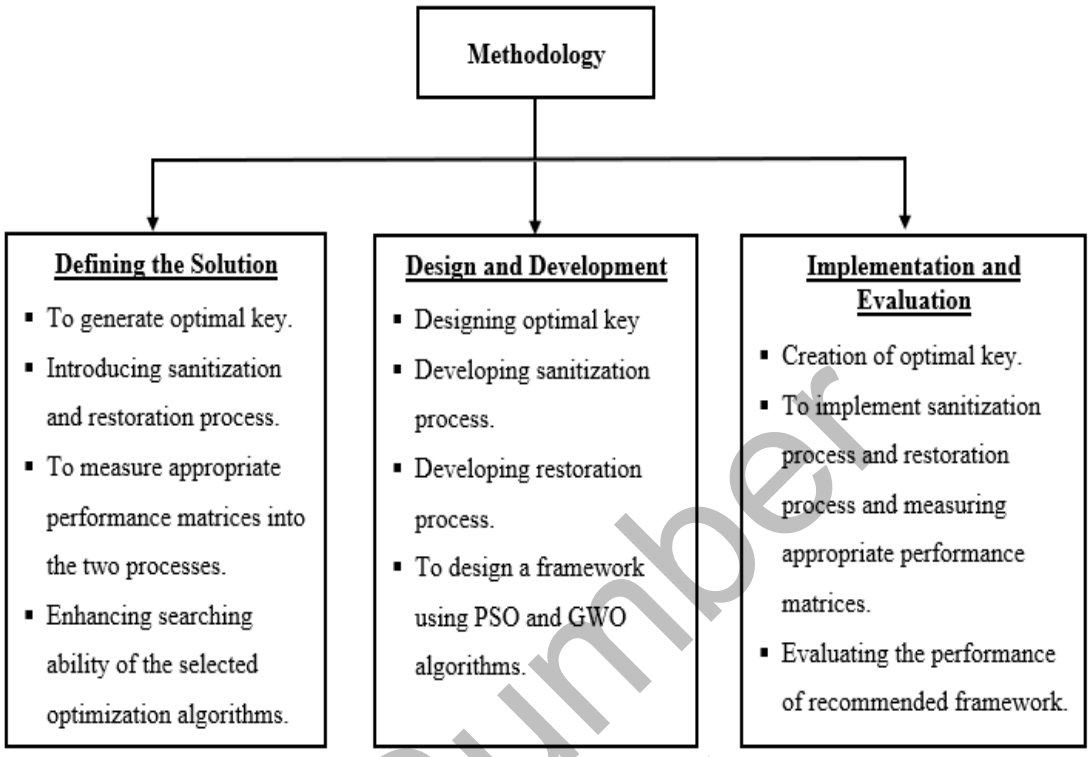



Figure 3.2 Three key points of research methodology for this research

### 3.4.1 Defining the Solution

This section of the research methodology aims to define the solutions of the problem. The purpose of this research is to conceal the sensitive data for the privacy of data. After reviewing literatures critically, as well as earlier recommended techniques and algorithms, the study identified the problem statements, research questions and objectives. After stating the problem statement, objective is obvious from the problem, accordingly, defining the solution as well as conducting literature review critically, objectives were obtained for achieving to get the promising performance. To accomplish the goals of this section, publications from reputable databases such as Web of Science, Scopus, IEEE, Science Direct, Elsevier and Springer are thoroughly reviewed to identify the methodologies used in the literature and to determine the concerns regarding data privacy problems. The critical analysis of these articles facilitates to find out problems and gives the directions to define solutions for obtaining objectives. A comprehensive literature review can be found in Chapter II.

### 3.4.2 Designing and Development

This design and development section presents the entire design and development of the recommended framework. After finding out the problems, such as generating optimal key including the duration of updating key value during the key generation stage, the allocation of key length, number of the key ranges, measuring appropriate performance matrices which have significant effects on sanitization and restoration processes for sanitizing and restoring data and convergence issues, like convergence to a point, slower convergence, not enough local searching ability of optimization algorithms etc. through conducting literature review, the design and development of a framework was suggested to fulfil the objectives. Main components of the recommended framework, for instances, are autism dataset as original data, support vector machine (SVM), sanitization process, restoration process, sanitization key, PSO and GWO algorithms and so on are shown in Figure 3.3.

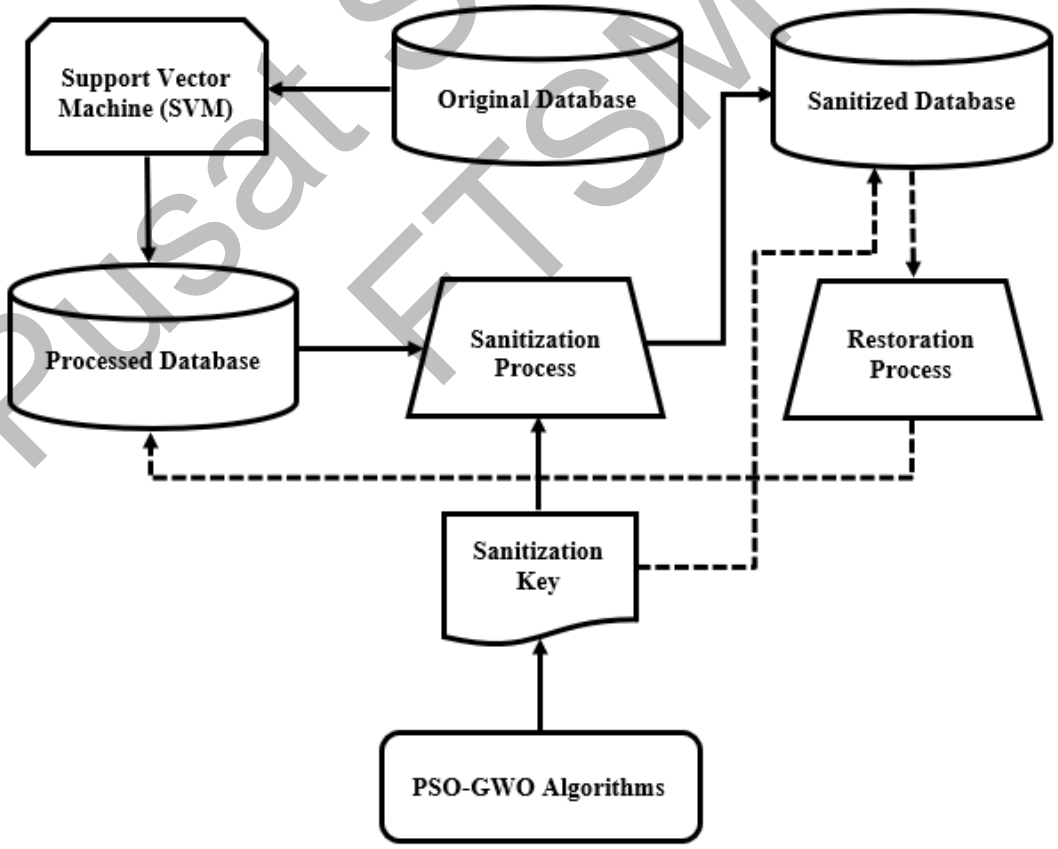



Figure 3.3 Overall architecture for data privacy preserving framework

Specifically, the different types of components for the architecture are summarized as follows:

1. Original Database;
2. Machine Learning;
3. Processed Database;
4. PSO-GWO Algorithms;
5. Sanitization Key;
6. Sanitization Process;
7. Sanitized Database;
8. Restoration Process.

The above components of the proposed framework are described broadly as below:

Original Database: The initial database which is selected for research in the thesis is known as original database. This original database is a raw database that is not prepared to be utilized, because some values can be missing, irregular, irrelevant or null. So, the original data need to be pre-processed. Here the four types of autism datasets are 24-months autism dataset, 30-months autism dataset, 36-months autism datasets, and 48-months autism datasets.

Support Vector Machine (SVM): Support Vector Machine is employed to the original datasets in order to get the processed data in this thesis. Using SVM, the datasets have been transformed into understandable format. It eliminates some descriptive questions which had no numerical values. It also managed the missing and duplicate values. By applying the SVM algorithm, all of the information is presented, such as total number of attributes, instances, data types, number of missing values, maximum and minimum values, error values, etc., easily and quickly. If any anomaly is occurred in the datasets, it can be solved easily.

Processed Database: Processed database comes from the original database by using a SVM algorithm. This database is ready to use because it is already transformed

to an understandable and desired format which have relevant data, not null values, or missing values.

PSO-GWO Algorithms: There are two optimization algorithms, such as particle swarm optimization and grey wolf optimization algorithm that are employed in this recommended framework. In the framework, the characteristics of PSO have been incorporated into GWO for enhancing the capability of convergence as well as local searching ability whose main purpose is to yield the best key.

Sanitization Key: The key which is utilized in hiding the data is known as sanitization key and it is produced from this framework using PSO-GWO algorithms. The key generation process including key encoding, key transformation is discussed broadly in the key generation section in chapter IV. An algorithm producing optimal key as well as how to be used that key in the sanitization process are also illustrated in the chapter IV. Moreover, the sanitization key is also utilized in the restoration process are shown broadly in chapter V.

Sanitization Process: The sanitization process is a process where processed data is being hidden by the presence of the sanitization key. This sanitization process has been illustrated in section 4.2 in chapter IV where different techniques, such as reconstruction of key matrix, khatri-rao process, binarization, XOR operation, and so on, have been applied.

Sanitized Database: The sanitized database is a database which is secured and protected and produced from processed database through the sanitization process. This sanitized database also be utilized to archive the processed data in the presence of sanitization key through restoration process.

Restoration Process: The restoration process is a technique where authorized person can get the processed data again by the presence of the sanitized database and sanitization key. The restoration process including decoding process and a restoration algorithm are discussed broadly in chapter V.

In this framework, the line arrow represents the sanitization process, which is the focus of this study, and the dash arrow denotes the restoration process. As an original data, I utilized autism datasets, for example, 24-months autism data, 30-months autism, 36-months autism data, 48-months autism data. Details of these datasets are illustrated in section 3.5.

### 3.4.3 Implementation and Evaluation

The research aim is to come up with a potential solution or remedy to the issues mentioned in problem statement. The significant issues that are pointed out in the problem statement section are generating inappropriate key including not mentioning the duration of updating key value during the key generation stage, the allocation of undefined key length, number of the key ranges, measuring not exact performance matrices which have significant effects on sanitization and restoration processes for sanitizing and restoring data and convergence issues, like slower convergence, not enough local searching ability of optimization algorithms etc. There are no proper existing technologies to definite resolution to the challenges in terms of privacy of medical data. Regarding those problems to address, the yielding of optimal keys is performed initially by using the hybridized characteristics of meta-heuristics algorithms. In this section, the study implemented a privacy preserving framework where the key is produced by this recommended Enhanced Combined PSO-GWO framework. There are also two processes, such as sanitization process and restoration process and this research chose appropriate fitness function to measure the performance for sanitization and restoration process regarding privacy of data. The research performed simulations using Python programming language taking the various types of autism datasets and compared the result with other traditional algorithms, such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Crow Search Algorithm (CSA), Differential Evolution (DE), and Adaptive Awareness Probability-based CSA (AAP-CSA) in order to establish the ideal solution and achieve enhanced competitive outputs. So, what implementations achieved in this thesis are in brief that, this work addressed those critical issues by forming the optimal key in the proper way, the sanitization process, and the restoration process with logical performance matrices related exactly to the privacy of medical sensitive data to provide the solutions for the

objectives of the thesis while the recommended framework was utilized. Noted that, the study illustrated broadly overall components of the proposed framework in the above section in this Chapter, whereas the generating optimal key, sanitization process and restoration process have been discussed in Chapter IV and Chapter V respectively.

However, the architecture of the proposed framework implementing by this way ensures the privacy of autism data and achieves the expected performances.

### 3.5 AUTISM DATASETS

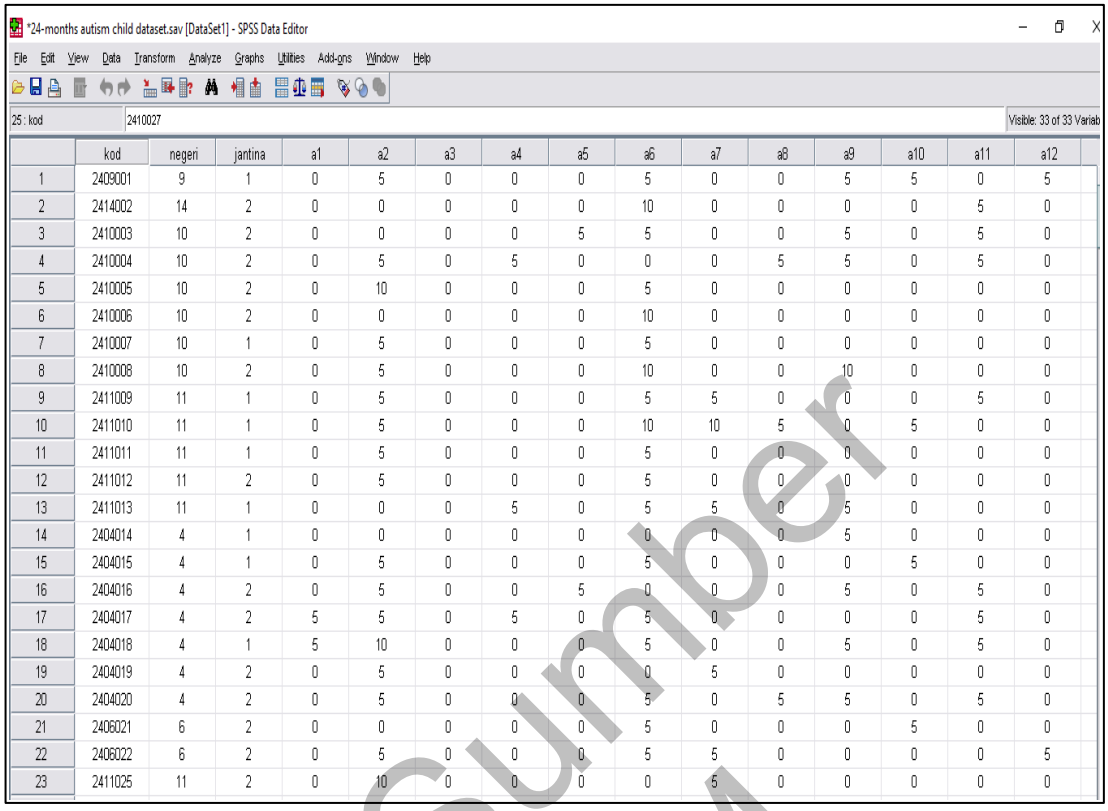
For the data privacy, this recommended framework has considered the medical health data, for example, autism datasets. There are different categories of data related to autism: screening, clinical (health) and nonclinical (non-health) (Thabtah 2019). Screening data are the data which are very sensitive data to diagnose whether individual is autistic or not. Broadly, clinical data is a medical data which is involved into health-related detail consisted of patient care regularly or a clinical trial program partly. Electronic health record is the best form of clinical data which is the digital version of medical information as well as their historical explanation in detail of the patients. Moreover, this electronic versions of record-keeping can ensure the efficiency of coordination and sharing information between sectors such as health, education, and social care (Fein & Helt 2017; The Lancet Neurology 2017) For example, it can be a general numerical data, such as vital signs as heart rate, respiration rate, and temperature, or diagnostic-related information, such as laboratory test results from blood tests, genetic testing, culture findings, and so on. It may also consist with images like as x-rays, as well as treatment information such as if the individual is taking any medications and, if so, how much or what dose they are taking and how often they are taking it, etc. Other form of medical data is non-clinical (non-health) data, the administrative data, focusing on record-keeping surrounding a service, for example, the information about hospital discharge. This might be included in the electronic health record as well. Others are claims data, which is the information about insurance claims; patient/disease registries, which are other platforms that assist in the collection and tracking of clinical data for specific patient groups; health surveys, which may assist to analyse or tally statistics such as the most frequent chronic illnesses a country

encounters; and clinical trial data, which is clinical information obtained via clinical research operations.

However, there are different aged autism datasets that have been employed for the simulations in this thesis, such as 24-months autism data, 30-months autism data, 36-months autism data, and 48-months autism data. These datasets were also utilized by Mohamed, S et al (Mohamed & Toran 2017) for their research regarding autism. The four different aged autism data are employed for the purposes of more comparison and better performing results in this thesis. The datasets are illustrated in the following:

### 3.5.1 24-months Autism Child Dataset

The 24-months autism dataset is collected from the faculty of Education, Universiti Kebangsaan Malaysia. This dataset consisted of 26 attributes and 209 instances which is depicted in Figure 3.4. The dataset has 26 questions, shown in Appendix A. Every question has 3 scoring options, such as z = 0, v = 5, and x = 10. This dataset also has 3 more questions asking for explanations only, not for scoring. In this case, cut-off score is 71.

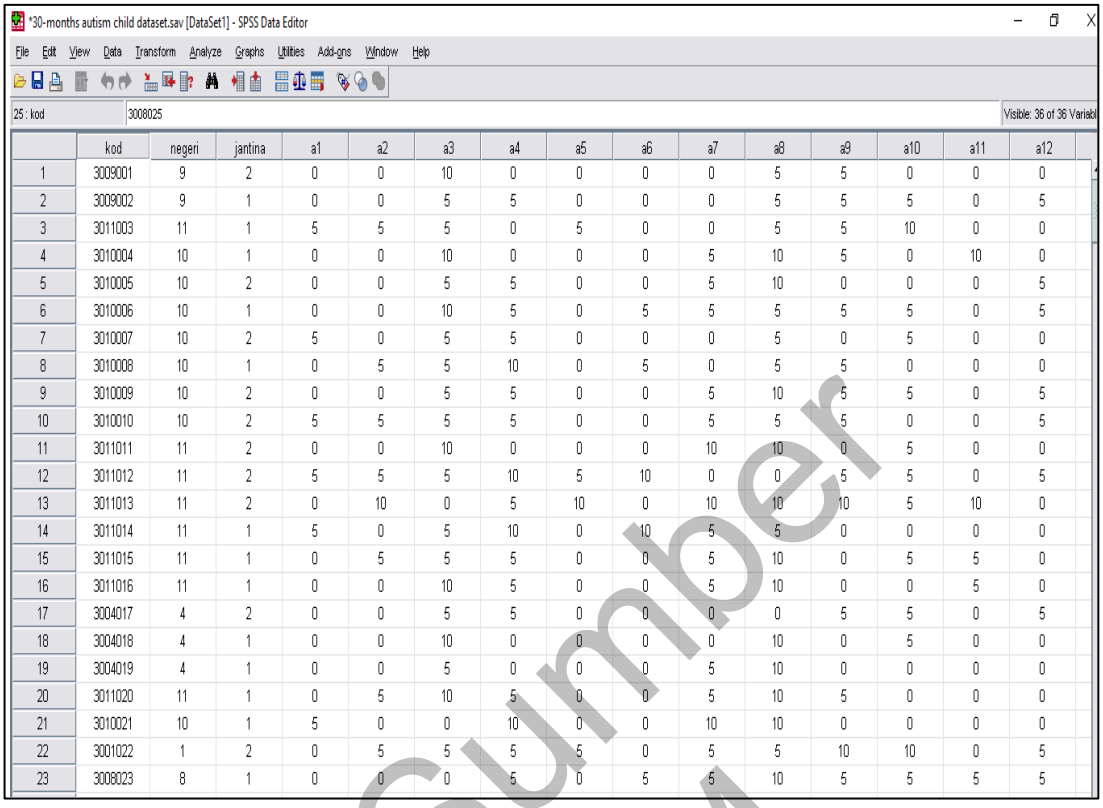



| | kod | negeri | jantina | a1 | a2 | a3 | a4 | a5 | a6 | a7 | a8 | a9 | a10 | a11 | a12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2409001 | 9 | 1 | 0 | 5 | 0 | 0 | 0 | 5 | 0 | 0 | 5 | 5 | 0 | 5 |
| 2 | 2414002 | 14 | 2 | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 5 | 0 |
| 3 | 2410003 | 10 | 2 | 0 | 0 | 0 | 0 | 5 | 5 | 0 | 0 | 5 | 0 | 5 | 0 |
| 4 | 2410004 | 10 | 2 | 0 | 5 | 0 | 5 | 0 | 0 | 0 | 5 | 5 | 0 | 5 | 0 |
| 5 | 2410005 | 10 | 2 | 0 | 10 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 2410006 | 10 | 2 | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 2410007 | 10 | 1 | 0 | 5 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 2410008 | 10 | 2 | 0 | 5 | 0 | 0 | 0 | 10 | 0 | 0 | 10 | 0 | 0 | 0 |
| 9 | 2411009 | 11 | 1 | 0 | 5 | 0 | 0 | 0 | 5 | 5 | 0 | 0 | 0 | 5 | 0 |
| 10 | 2411010 | 11 | 1 | 0 | 5 | 0 | 0 | 0 | 10 | 10 | 5 | 0 | 5 | 0 | 0 |
| 11 | 2411011 | 11 | 1 | 0 | 5 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 2411012 | 11 | 2 | 0 | 5 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | 2411013 | 11 | 1 | 0 | 0 | 0 | 5 | 0 | 5 | 5 | 0 | 5 | 0 | 0 | 0 |
| 14 | 2404014 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 |
| 15 | 2404015 | 4 | 1 | 0 | 5 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 5 | 0 | 0 |
| 16 | 2404016 | 4 | 2 | 0 | 5 | 0 | 0 | 5 | 0 | 0 | 0 | 5 | 0 | 5 | 0 |
| 17 | 2404017 | 4 | 2 | 5 | 5 | 0 | 5 | 0 | 5 | 0 | 0 | 0 | 0 | 5 | 0 |
| 18 | 2404018 | 4 | 1 | 5 | 10 | 0 | 0 | 0 | 5 | 0 | 0 | 5 | 0 | 5 | 0 |
| 19 | 2404019 | 4 | 2 | 0 | 5 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 |
| 20 | 2404020 | 4 | 2 | 0 | 5 | 0 | 0 | 0 | 5 | 0 | 5 | 5 | 0 | 5 | 0 |
| 21 | 2406021 | 6 | 2 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 5 | 0 | 0 |
| 22 | 2406022 | 6 | 2 | 0 | 5 | 0 | 0 | 0 | 5 | 5 | 0 | 0 | 0 | 0 | 5 |
| 23 | 2411025 | 11 | 2 | 0 | 10 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 |

Figure 3.4 24 months autism child dataset

### 3.5.2 30-months Autism Child Dataset

The 30-months autism dataset is also collected from the faculty of Education, Universiti Kebangsaan Malaysia. This dataset consisted of 29 attributes and 209 instances, shown in Figure 3.5. The dataset has 29 questions that is mentioned Appendix A. Here, every question has also 3 scoring options, such as z = 0, v = 5, and x = 10. Similarly, the dataset has 3 additional questions which are for explanations only, not for scoring values. Cut-off score is 95 for this dataset.

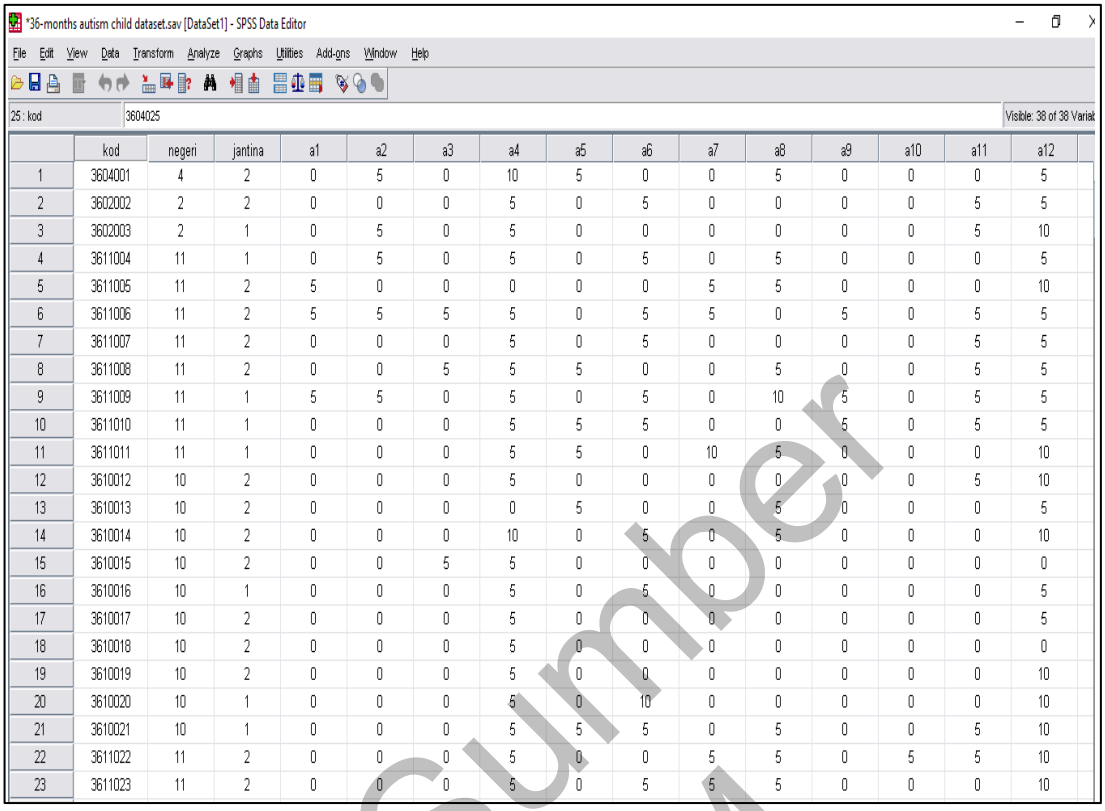



| | kod | negeri | jantina | a1 | a2 | a3 | a4 | a5 | a6 | a7 | a8 | a9 | a10 | a11 | a12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3009001 | 9 | 2 | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 5 | 5 | 0 | 0 | 0 |
| 2 | 3009002 | 9 | 1 | 0 | 0 | 5 | 5 | 0 | 0 | 0 | 5 | 5 | 5 | 0 | 5 |
| 3 | 3011003 | 11 | 1 | 5 | 5 | 5 | 0 | 5 | 0 | 0 | 5 | 5 | 10 | 0 | 0 |
| 4 | 3010004 | 10 | 1 | 0 | 0 | 10 | 0 | 0 | 0 | 5 | 10 | 5 | 0 | 10 | 0 |
| 5 | 3010005 | 10 | 2 | 0 | 0 | 5 | 5 | 0 | 0 | 5 | 10 | 0 | 0 | 0 | 5 |
| 6 | 3010006 | 10 | 1 | 0 | 0 | 10 | 5 | 0 | 5 | 5 | 5 | 5 | 5 | 0 | 5 |
| 7 | 3010007 | 10 | 2 | 5 | 0 | 5 | 5 | 0 | 0 | 0 | 5 | 0 | 5 | 0 | 0 |
| 8 | 3010008 | 10 | 1 | 0 | 5 | 5 | 10 | 0 | 5 | 0 | 5 | 5 | 0 | 0 | 0 |
| 9 | 3010009 | 10 | 2 | 0 | 0 | 5 | 5 | 0 | 0 | 5 | 10 | 5 | 5 | 0 | 5 |
| 10 | 3010010 | 10 | 2 | 5 | 5 | 5 | 5 | 0 | 0 | 5 | 5 | 5 | 0 | 0 | 5 |
| 11 | 3011011 | 11 | 2 | 0 | 0 | 10 | 0 | 0 | 0 | 10 | 10 | 0 | 5 | 0 | 0 |
| 12 | 3011012 | 11 | 2 | 5 | 5 | 5 | 10 | 5 | 10 | 0 | 0 | 5 | 5 | 0 | 5 |
| 13 | 3011013 | 11 | 2 | 0 | 10 | 0 | 5 | 10 | 0 | 10 | 10 | 10 | 5 | 10 | 0 |
| 14 | 3011014 | 11 | 1 | 5 | 0 | 5 | 10 | 0 | 10 | 5 | 5 | 0 | 0 | 0 | 0 |
| 15 | 3011015 | 11 | 1 | 0 | 5 | 5 | 5 | 0 | 0 | 5 | 10 | 0 | 5 | 5 | 0 |
| 16 | 3011016 | 11 | 1 | 0 | 0 | 10 | 5 | 0 | 0 | 5 | 10 | 0 | 0 | 5 | 0 |
| 17 | 3004017 | 4 | 2 | 0 | 0 | 5 | 5 | 0 | 0 | 0 | 0 | 5 | 5 | 0 | 5 |
| 18 | 3004018 | 4 | 1 | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 10 | 0 | 5 | 0 | 0 |
| 19 | 3004019 | 4 | 1 | 0 | 0 | 5 | 0 | 0 | 0 | 5 | 10 | 0 | 0 | 0 | 0 |
| 20 | 3011020 | 11 | 1 | 0 | 5 | 10 | 5 | 0 | 0 | 5 | 10 | 5 | 0 | 0 | 0 |
| 21 | 3010021 | 10 | 1 | 5 | 0 | 0 | 10 | 0 | 0 | 10 | 10 | 0 | 0 | 0 | 0 |
| 22 | 3001022 | 1 | 2 | 0 | 5 | 5 | 5 | 5 | 0 | 5 | 5 | 10 | 10 | 0 | 5 |
| 23 | 3008023 | 8 | 1 | 0 | 0 | 0 | 5 | 0 | 5 | 5 | 10 | 5 | 5 | 5 | 5 |

Figure 3.5 30 months autism child dataset

### 3.5.3 36-months Autism Child Dataset

The dataset includes 31 attributes and 234 instances that partly shown in Figure 3.6. The dataset consisted of 31 questions and illustrated in Appendix A. Here, every question has 3 scoring options, such as z = 0, v = 5, and x = 10 as earlier dataset. The dataset also has 3 additional questions for explanations, not for scoring. Noted that, for the dataset, cut-off score is 100.

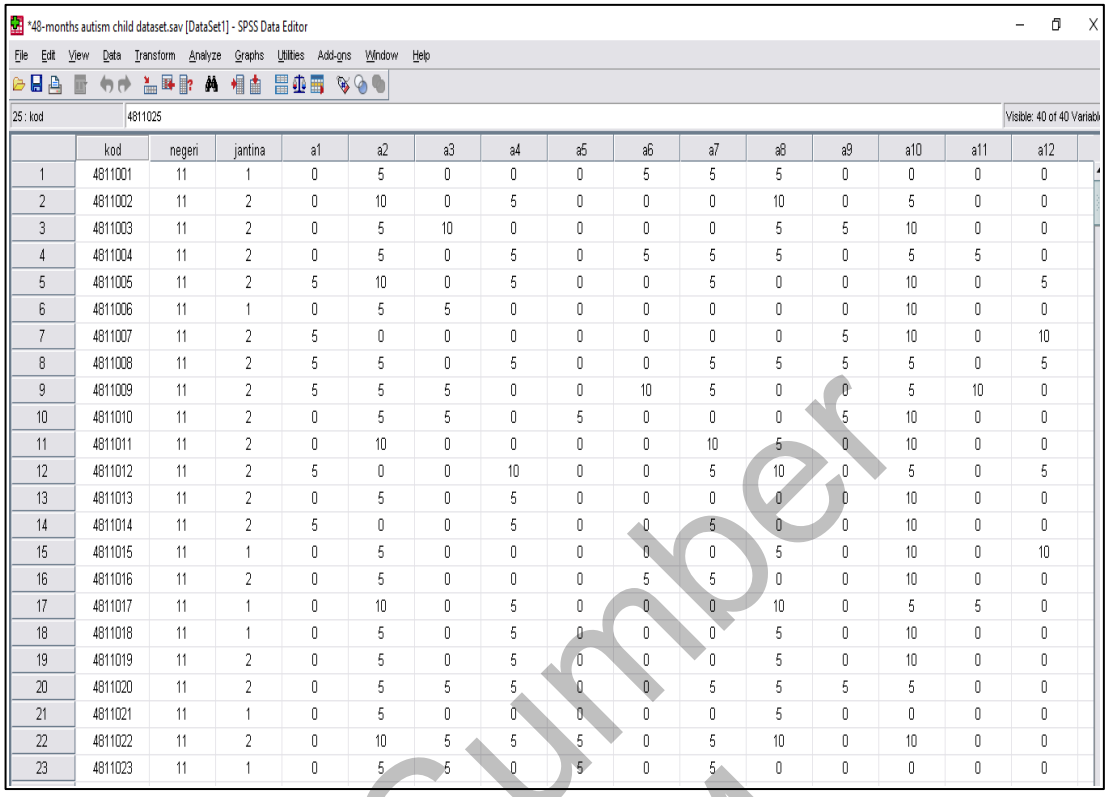



| | kod | negeri | jantina | a1 | a2 | a3 | a4 | a5 | a6 | a7 | a8 | a9 | a10 | a11 | a12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3604001 | 4 | 2 | 0 | 5 | 0 | 10 | 5 | 0 | 0 | 5 | 0 | 0 | 0 | 5 |
| 2 | 3602002 | 2 | 2 | 0 | 0 | 0 | 5 | 0 | 5 | 0 | 0 | 0 | 0 | 5 | 5 |
| 3 | 3602003 | 2 | 1 | 0 | 5 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 10 |
| 4 | 3611004 | 11 | 1 | 0 | 5 | 0 | 5 | 0 | 5 | 0 | 5 | 0 | 0 | 0 | 5 |
| 5 | 3611005 | 11 | 2 | 5 | 0 | 0 | 0 | 0 | 0 | 5 | 5 | 0 | 0 | 0 | 10 |
| 6 | 3611006 | 11 | 2 | 5 | 5 | 5 | 5 | 0 | 5 | 5 | 0 | 5 | 0 | 5 | 5 |
| 7 | 3611007 | 11 | 2 | 0 | 0 | 0 | 5 | 0 | 5 | 0 | 0 | 0 | 0 | 5 | 5 |
| 8 | 3611008 | 11 | 2 | 0 | 0 | 5 | 5 | 5 | 0 | 0 | 5 | 0 | 0 | 5 | 5 |
| 9 | 3611009 | 11 | 1 | 5 | 5 | 0 | 5 | 0 | 5 | 0 | 10 | 5 | 0 | 5 | 5 |
| 10 | 3611010 | 11 | 1 | 0 | 0 | 0 | 5 | 5 | 5 | 0 | 0 | 5 | 0 | 5 | 5 |
| 11 | 3611011 | 11 | 1 | 0 | 0 | 0 | 5 | 5 | 0 | 10 | 5 | 0 | 0 | 0 | 10 |
| 12 | 3610012 | 10 | 2 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 10 |
| 13 | 3610013 | 10 | 2 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 5 | 0 | 0 | 0 | 5 |
| 14 | 3610014 | 10 | 2 | 0 | 0 | 0 | 10 | 0 | 5 | 0 | 5 | 0 | 0 | 0 | 10 |
| 15 | 3610015 | 10 | 2 | 0 | 0 | 5 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | 3610016 | 10 | 1 | 0 | 0 | 0 | 5 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 5 |
| 17 | 3610017 | 10 | 2 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 18 | 3610018 | 10 | 2 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | 3610019 | 10 | 2 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| 20 | 3610020 | 10 | 1 | 0 | 0 | 0 | 5 | 0 | 10 | 0 | 0 | 0 | 0 | 0 | 10 |
| 21 | 3610021 | 10 | 1 | 0 | 0 | 0 | 5 | 5 | 5 | 0 | 5 | 0 | 0 | 5 | 10 |
| 22 | 3611022 | 11 | 2 | 0 | 0 | 0 | 5 | 0 | 0 | 5 | 5 | 0 | 5 | 5 | 10 |
| 23 | 3611023 | 11 | 2 | 0 | 0 | 0 | 5 | 0 | 5 | 5 | 5 | 0 | 0 | 0 | 10 |

Figure 3.6 36 months autism child dataset

### 3.5.4 48-months Autism Child Dataset

This dataset comprised of 33 attributes and 302 instances which revealed in Figure 3.7. The dataset includes 33 questions which is attached in Appendix A. Here, every question has 3 scoring options, such as z = 0, v = 5, and x = 10 like other dataset. The dataset has 3 additional questions as well, which have no scoring values. In this case, cut-off score is 105.




| | kod | negeri | jantina | a1 | a2 | a3 | a4 | a5 | a6 | a7 | a8 | a9 | a10 | a11 | a12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4811001 | 11 | 1 | 0 | 5 | 0 | 0 | 0 | 5 | 5 | 5 | 0 | 0 | 0 | 0 |
| 2 | 4811002 | 11 | 2 | 0 | 10 | 0 | 5 | 0 | 0 | 0 | 10 | 0 | 5 | 0 | 0 |
| 3 | 4811003 | 11 | 2 | 0 | 5 | 10 | 0 | 0 | 0 | 0 | 5 | 5 | 10 | 0 | 0 |
| 4 | 4811004 | 11 | 2 | 0 | 5 | 0 | 5 | 0 | 5 | 5 | 5 | 0 | 5 | 5 | 0 |
| 5 | 4811005 | 11 | 2 | 5 | 10 | 0 | 5 | 0 | 0 | 5 | 0 | 0 | 10 | 0 | 5 |
| 6 | 4811006 | 11 | 1 | 0 | 5 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 0 |
| 7 | 4811007 | 11 | 2 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 10 | 0 | 10 |
| 8 | 4811008 | 11 | 2 | 5 | 5 | 0 | 5 | 0 | 0 | 5 | 5 | 5 | 5 | 0 | 5 |
| 9 | 4811009 | 11 | 2 | 5 | 5 | 5 | 0 | 0 | 10 | 5 | 0 | 0 | 5 | 10 | 0 |
| 10 | 4811010 | 11 | 2 | 0 | 5 | 5 | 0 | 5 | 0 | 0 | 0 | 5 | 10 | 0 | 0 |
| 11 | 4811011 | 11 | 2 | 0 | 10 | 0 | 0 | 0 | 0 | 10 | 5 | 0 | 10 | 0 | 0 |
| 12 | 4811012 | 11 | 2 | 5 | 0 | 0 | 10 | 0 | 0 | 5 | 10 | 0 | 5 | 0 | 5 |
| 13 | 4811013 | 11 | 2 | 0 | 5 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 0 |
| 14 | 4811014 | 11 | 2 | 5 | 0 | 0 | 5 | 0 | 0 | 5 | 0 | 0 | 10 | 0 | 0 |
| 15 | 4811015 | 11 | 1 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 10 | 0 | 10 |
| 16 | 4811016 | 11 | 2 | 0 | 5 | 0 | 0 | 0 | 5 | 5 | 0 | 0 | 10 | 0 | 0 |
| 17 | 4811017 | 11 | 1 | 0 | 10 | 0 | 5 | 0 | 0 | 0 | 10 | 0 | 5 | 5 | 0 |
| 18 | 4811018 | 11 | 1 | 0 | 5 | 0 | 5 | 0 | 0 | 0 | 5 | 0 | 10 | 0 | 0 |
| 19 | 4811019 | 11 | 2 | 0 | 5 | 0 | 5 | 0 | 0 | 0 | 5 | 0 | 10 | 0 | 0 |
| 20 | 4811020 | 11 | 2 | 0 | 5 | 5 | 5 | 0 | 0 | 5 | 5 | 5 | 5 | 0 | 0 |
| 21 | 4811021 | 11 | 1 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 |
| 22 | 4811022 | 11 | 2 | 0 | 10 | 5 | 5 | 5 | 0 | 5 | 10 | 0 | 10 | 0 | 0 |
| 23 | 4811023 | 11 | 1 | 0 | 5 | 5 | 0 | 5 | 0 | 5 | 0 | 0 | 0 | 0 | 0 |

Figure 3.7 48 months autism child dataset

### 3.5.5 Autism Data Validation

Datasets for above four ASQ: SE (M) (e.g., 24-, 30-, 36-, 48- months autism children data) were validated by 8 experts in early childhood development measuring behaviour, self-control, compliance, communication, self-adjustment, autonomy and interaction (Mohamed & Toran 2017). And the data are standard by maintaining some important criteria. They are discussed at below:

The methodology is given in Table 3.1.

Table 3.1 Methodology of data collection

| Research design | Location | Sample | Data collection procedures | Data analysis procedures |
|---|---|---|---|---|
| Mixed methods (explanatory) | Peninsular Malaysia | 954 | Focus group discussion & questionnaires | Qualitative (descriptive analysis), Statistical Package for Social Sciences (SPSS) and Receiver Operating Characteristic (ROC) |

Every data type has similar content and format, so the internal consistency of the datasets is shown in Table 3.2.

Table 3.2 Internal consistency of ASQ:SE (M)

| ASQ:SE (M) | *n* | Alpha |
|---|---|---|
| 24 months | 209 | 0.67 |
| 30 months | 209 | 0.73 |
| 36 months | 234 | 0.84 |
| 48 months | 302 | 0.80 |

For each type of dataset has different cut-off value, sensitivity, and specificity and illustrated in Table 3.3.

Table 3.3 Cut-off, sensitivity, and specificity of ASQ:SE (M)

| ASQ:SE (M) | Cut-off score | Sensitivity | Specificity |
|---|---|---|---|
| 24 months | 71 | 83 | 98 |
| 30 months | 95 | 71 | 98 |
| 36 months | 100 | 67 | 94 |
| 48 months | 105 | 73 | 95 |

And concurrent validity for the data is revealed in Table 3.4.

Table 3.4 Concurrent validity of ASQ:SE (M)

| ASQ:SE (M) | *n* | Agreement (%) |
|---|---|---|
| 24 months | 50 | 90.0 |
| 30 months | 50 | 95.0 |
| 36 months | 50 | 96.0 |
| 48 months | 50 | 98.0 |



## 3.6 WEKA SOFTWARE

Waikato Environment for Knowledge Analysis (WEKA) is a popular data mining software which was developed at the University of Waikato, New Zealand. This software is one of the most commonly used in data mining systems. It is started through the necessary requirements for integrated desk work, which enables researchers to access sophisticated methods in machine learning (ML) easily. At the time of the

establishment of a project in 1992, learning algorithms were available in many languages, various platforms, and different kinds of data formats, which made them very challenging, so WEKA was expected to be a tool that would provide not only learning algorithms but also a framework where researchers could apply new algorithms without having to deal with data manipulation and support for infrastructure schema. The original version of WEKA was established in the domain of agriculture. Though the fundamental version of WEKA was initially developed as a tool for analysing data from the agricultural domain, nowadays it is greatly utilised in various datasets concerning issues in the sectors of engineering, bioinformatics, business, medical, statistics, etc. There are numerous learning methods, techniques, algorithms, and functions in WEKA for various purposes, such as processing, clustering, association, regression, classification, visualization, etc. of data. Users utilise these algorithms and functions separately or collectively as needed.

## 3.7 PYTHON PROGRAMMING LANGUAGE

Python is an open source, more scalable, object-oriented, scripting programming language which is very useful in cybersecurity. This language was invented by Guido van Rossum in the early 90s in the Netherlands. Though there are some distributions of Python, in this regard, I used the Python 3.8 distribution for simulations and analysis. Python is an incredibly helpful programming language for cybersecurity experts since it can handle a wide range of activities such as host discovery, network scanning, port scanning, malware analysis, penetration testing, transmission of packets, and accessing servers. It is also one of the most sophisticated programming languages, which means a high degree of web readability. Consequently, for its many advantages and robustness, it is used by some of the world's most well-known digital corporations, for instance, NASA, Google, and Reddit, among others.

## 3.8 SUMMARY

This chapter illustrates the research methodology for this research work. The chapter describes the phases of working progress during the research and three key sections for this research in details. Three key sections, namely, defining solutions, designing and developments, implementation and evaluation for this research are illustrated broadly

and sequentially as the research methodology process. The research questions and the research problems are identified depending on the literature review earlier chapter and the evaluation, finally, is come true through the proper methodology explained in this chapter. This chapter explains the procedures of specifying the problems, designing and developing the techniques and finally implementation and evaluation of the framework in order to fulfil the desired outcome for the data privacy regarding autism.

# CHAPTER IV

# THE ENHANCEMENT OF THE OPTIMIZATION KEY AND SANITIZATION PROCESS FOR AUTISM DATA

## 4.1 INTRODUCTION

An optimal key is a sanitization key that is employed on datasets to hide the data. This key is utilized in a process by which processed datasets are not accessible to others. The process is known as the sanitization process, and the non-accessible datasets are sanitised data. According to the literature review, the following issues should be addressed regarding the optimal key and sanitization process for enhancement of the privacy of data:

- How long will the key value be updated during the key generation stage?
- The key length will be allocated based on which value?
- Defining values of the parameters.
- A number of the key ranges.

This study generates optimal key firstly, and after that, a process named the data sanitization process has been applied for the better privacy of autism datasets. The optimal key is created through the enhanced combined PSO-GWO framework without compromising the above issues. Data sanitization is a procedure where sensitive information is disguised to make test and development databases (Edgar 2004). This can be done by overwriting it with similar types of false data but making it look realistic. Though there are various data sanitization techniques, such as encryption and decryption, gibberish generation, number variance, shuffling records, substitution, masking data, etc., the study presented here has applied the optimal key and sensitization process.

However, the contribution of this study in this section can be summarized as below:

- First, to generate an optimal key by considering the above issues.
- Secondly, to enhance a data sanitization process where an optimal key is used for the privacy of ASD datasets.
- Finally, comparing the performances achieved by this sanitization process, including optimal key, with the performances of other existing privacy frameworks.

## 4.2 ARCHITECTURE FOR SANITIZING AUTISM SENSITIVE DATA

Autism sensitive data protection has been implemented by data sanitization technique. The following Figure 4.1 is the architecture of sanitizing autism sensitive data, which ensures the privacy of autism data and maintains the expected results. Noted that, from the main overall architecture of the framework in Figure 3.3, the line arrow represents the sanitization process which is one of the focuses of this research, and the dash arrow denotes the restoration process that are another objective of this research. As an objective of this research work, the sanitization of autism sensitive data is illustrated in the following Figure 4.1 and the next subsequent figures in details.

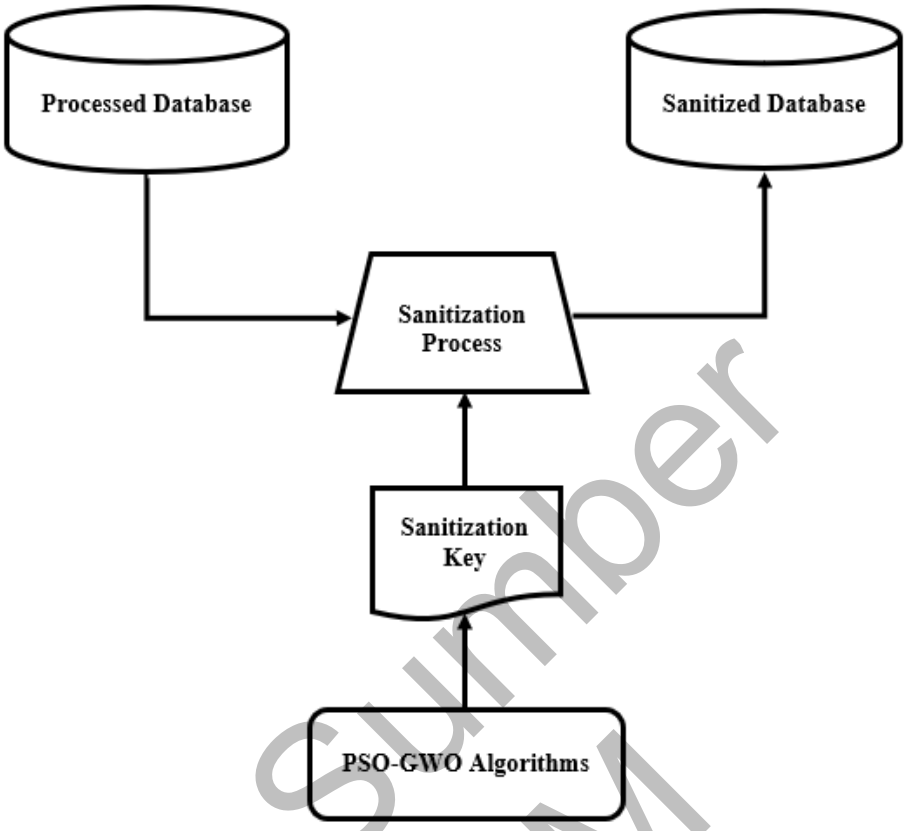



Figure 4.1 Architecture for sanitizing autism sensitive data.

This partial architecture of the main overall architecture for sanitization contains the following significant components as below:

1. Processed Database
2. PSO-GWO Algorithms
3. Sanitization Key
4. Sanitization Process
5. Sanitized Database.

The above components are illustrated in the design and development section in chapter III. In addition, the mathematical equations that have utilized symbols and operators are summarized in the following Table 4.1.

Table 4.1 List of mathematical symbols used in data sanitization process

| **Symbols** | **Descriptions** |
|---|---|
| $D$ | Processed (from original) database |
| $D'$ | Sanitization database |
| $K_1, K_2 \ldots K_N$ | Number of keys |
| $K_2$ | Pruned key matrix |
| $\oplus$ | XOR operator |
| + | Binary Summation |
| $\lfloor \ \rfloor$ | Floor function |
| $L_D$ | Sanitization key length |
| $\sqrt{L_D^C}$ | Key length |
| $T_1, T_2, \ldots T_5$ | Number of transactions |
| $T_{max}$ | Maximum transaction |
| $\otimes$ | Kronecker product |
| $C_1, C_2, C_3$ | Objective functions |
| $f_s$ | Frequency of sensitive itemset in sanitized data |
| $f_m$ | Frequency of sensitive itemset in original data |
| $f_{ns}$ | Frequency of non-sensitive itemset in sanitized data |
| $w_1, w_2, w_3$ | Impact of a particular cost function |
| $f$ | Fitness function |
| G | Minimum objective function |
| $\vec{M}$ | Location of the particle |
| $\vec{w}$ | Velocity of the particle |
| ω | User-defined behavioural parameter (an inertia weight) |
| $\vec{q}$ | Particle's previous best position (pbest position) |
| $\vec{f}$ | Particle's previous best position in the swarm (gbest position) |
| $r_1, r_2$ | Stochastic variables |
| $c_1, c_2$ | Acceleration constants |
| $u$ | Current iteration |
| $\vec{H}, \vec{E}$ | Coefficient vectors |

### 4.2.1 Data Sanitization

The procedure of the sanitization technique is illustrated in Figure 4.2 Here $D´$ , a sanitization database, is obtained accompanied by the sanitizing key generated from the initial database during the key generation process.

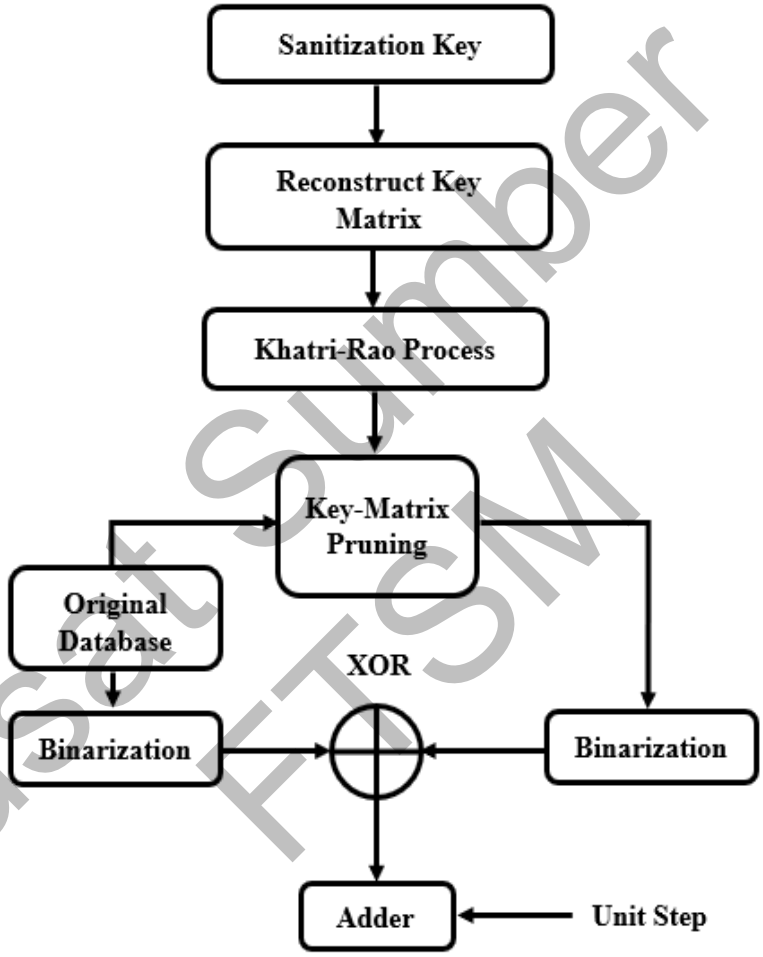



Figure 4.2 The architecture of the sanitization process.

The resulting key matrix, $K_2$, and $D$ indicate the pruned key matrix and original database, respectively, that are binarized to fulfil the XOR function. Following this binary XOR operation, the chance of having '0' is high. Getting such zeros yield insignificant data elements. So, for avoiding such zeros, a unit value (one) is added where the + (plus) sign refers to the binary summation. Then a unit step input is summed up consequently as well as $D´$ is obtained as the following in Equation 4.1 (Ahamad et al. 2022),

$$D' = (K_2 \oplus D) + 1 \quad \ldots(4.1)$$

### 4.2.2 Key Generation

Figure 4.3 demonstrates the key generation process for sanitization purposes. The optimal key is created with the help of the Enhanced Combined PSO-GWO framework by setting the population of various keys indiscriminately. It is followed by the sanitization process step through that sanitized database is obtained. Specifically, Figure 4.3 illustrates the key generation process for data sanitization and the restoration process.

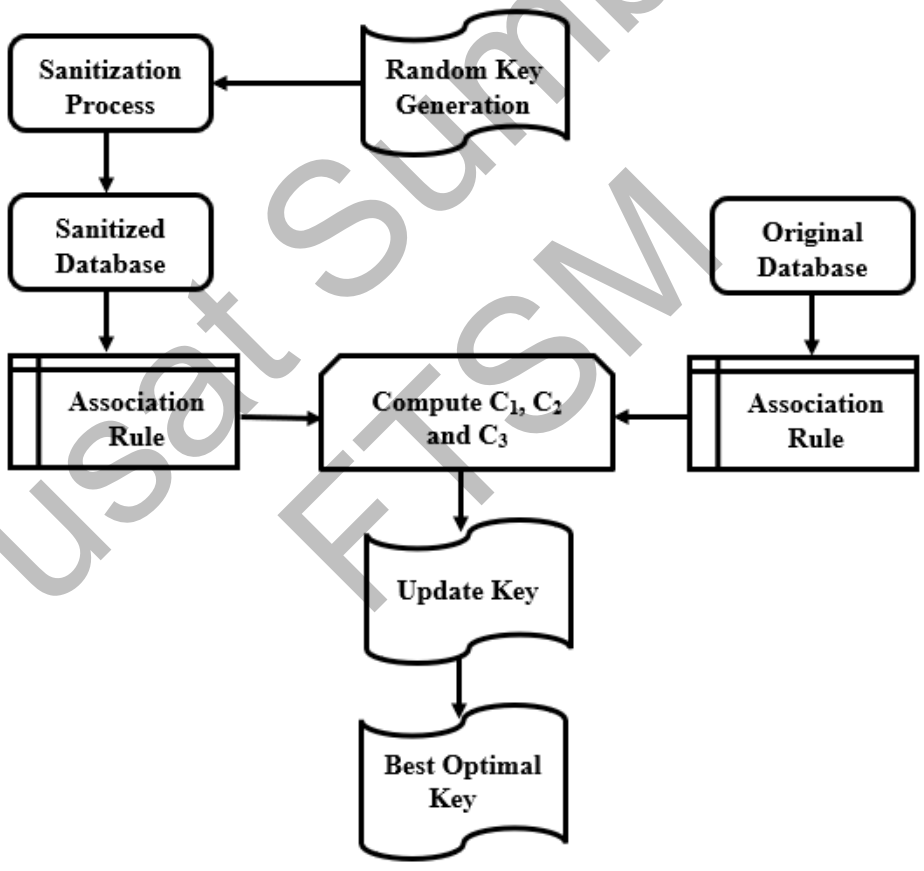



Figure 4.3 The architecture of the proposed key generation process

The Enhanced Combined PSO-GWO algorithm is used at the key update step for obtaining the better key and is performed depending on an iterative loop to get the better solution in the process. In the interim, the sanitized database is obtained through the sanitization process. Again, the original database acquires an association rule and

measure the objective functions, $C_1$, $C_2$, and $C_3$, respectively. Finally, the key value is updated continuously during this process until the highest termination measure is achieved and the best-desired solution is generated. For this data sanitization process, a key is created optimally by the Enhanced Combined PSO-GWO. The dimension of the chromosome is allotted depending on the value of $\sqrt{L\ _D^C}$ . The value fixes the elements, $\lfloor 0, \sqrt{\max(D)} \rfloor$, whereas *D* refers to the initial database.

### 4.2.3 Procedure of Proposed Optimal Key Extraction in Sanitization Process

- Key Encoding

The usage of keys, *K* for the procedure of sanitization depends on the encoding of the Enhanced Combined PSO-GWO algorithm. The optimization of the number of keys ranges from key $K_1$ to key $K_N$ is controlled by using an Enhanced Combined PSO-GWO algorithm, and as a result, the optimal key is obtained. The length of the key is allotted as $\sqrt{L\ _D^C}$ in this case. Usually, the key length for sanitization is to be $L_D$. However, the key generation process needs $\sqrt{L\ _D^C}$ and the technique of key transformation forms a key of $L_D$ using Khatri–Rao product. A Kronecker product that is column-wise is known as the Khatri–Rao product (Freitas Jr et al. 2018).

- Key Transformation

Let's consider a database transaction in Table 4.2,

Table 4.2 Data transactions in the database

| Transactions | Data | | |
|---|---|---|---|
| $T_1$ | 1 | 2 | 0 |
| $T_2$ | 1 | 3 | 0 |
| $T_3$ | 2 | 3 | 4 |
| $T_4$ | 1 | 3 | 4 |
| $T_5$ | 3 | 4 | 0 |

The key K is converted by applying the Khatri-Rao product during the Key Transformation process phase. This operation is occurred on two matrices of arbitrary size as a block matrix and is denoted by the operator ⊗. From the beginning, *K* is mainly formed as $K_1$ with the dimension of the matrix, [ $\sqrt{L\frac{C}{D}} \times T_{\max}$]. The recommended technique of $K$ = 5, 0, 10, for illustration, performs row-wise duplication and produces the key matrix, $K_1$ with dimension [$\sqrt{L\frac{C}{D}} \times T_{\max}$] as revealed in Equation 4.2, wherein the row matrix depends on $\sqrt{L\frac{C}{D}}$ as well as the column matrix, is assigned depending on $T_{max}$.

So, the matric *K* with size, [$\sqrt{L\frac{C}{D}} \times T_{\max}$]

$$K_1 = \begin{bmatrix} 5 & 5 & 5 \\ 0 & 0 & 0 \\ 10 & 10 & 10 \end{bmatrix} \quad \ldots(4.2)$$

Similarly, by applying Khatri–Rao products like $K_1 \otimes K_1$, the key matrix, $K_2$, is achieved whose dimension is [ $L_D \times T_{\max}$]. Its sizes are trimmed regarding the initial database dimensions presented through Equation 4.3,

$$K_2 = \begin{bmatrix} 5 & 5 & 5 \\ 0 & 0 & 0 \\ 10 & 10 & 10 \end{bmatrix} \otimes \begin{bmatrix} 5 & 5 & 5 \\ 0 & 0 & 0 \\ 10 & 10 & 10 \end{bmatrix} \quad \ldots(4.3)$$

$K_1$ acts the key generation process depending on Khatri–Rao approach and produces a matrix of sizes same as the initial database, $K_2$ [ $L_D \times T_{\max}$]. Finally, the rule hiding method is encompassed to obtain the sanitized database, $D'$ by concealing the sensitive data. Besides, binarization is performed between the original database as well as the key matrix. Consequently, the rule hiding operation is applied to the binarized key matrix pruning, wherein XOR function takes place with initial binarized database accomplishing equivalent matrix sizes and adds up with unit value and produce the sanitized database, which is revealed in Equation 4.1, where $K_2$ implies pruned key matrix. Furthermore, earlier to sanitization of *D*, $D'$ achieved from the sanitization process raises both sensitive rules and association rules. In this way, Equation 4.1 is analysed depending on Khatri–Rao method and is reached by sanitized database $D'$.

- Assessment of Fitness Functions

The functions, $C_1$, $C_2$, and $C_3$, known as objective functions. $C_1$ is hiding failure rate, $C_2$ express the information loss rate, and $C_3$ denotes the degree of modification which are assessed through Equation 4.4 to Equation 4.6 after sensitive rules, and association rules of the original and sanitized database are generated in this work. In Equation 4.4 (Mewada et al. 2020), $f_s$ and $f_m$ refer to the frequency of sensitive itemset, whereas $f_s$ signifies in the case of sanitized data, and $f_m$ implies in respect of original data.

$$C_1 = \frac{fs}{fm} \quad \ldots(4.4)$$

Similarly, $f_{ns}$ represents the non-sensitive itemset frequency in reference to sanitized data shown in Equation 4.5 (Mewada et al. 2020),

$$C_2 = \frac{fns}{fm} \quad \ldots(4.5)$$

From Equation 4.6 (Ahamad et al. 2022), the Euclidean distance is achieved where $D$ is original data, and $D'$ is sanitized data.

$$C_3 = dist\,(D, D'), \quad \text{where}, dist \rightarrow Euclidean\ distance \quad \ldots(4.6)$$

Initially, the function is expressed as in Equation 4.7 below,

$$Max\ f(x) = \sum_{i=1,\ j=1}^{i=M,\ j=N} (w_j\, f_i) \qquad where, \{ f_i : \forall_i = M \} \quad \ldots(4.7)$$

$$\{ w_j : \forall_j = N \}$$

$$= w_1 x_1 + w_2 x_2 + w_3 x_3$$

$$= w_1 f_1 + w_2 f_2 + w_3 f_3 \qquad where, x = \{ f_1, f_2, f_3 \}$$

Here, x also implies the objective functions, $f_1$, $f_2$, and $f_3$ wherein $f_1$ refers information hiding rate, $f_2$ indicates information preservation rate, and $f_3$ denotes the degree of non-modification rate which all should be maximized. In addition, $w_1$, $w_2$, $w_3$ represent the particular cost function.

Now the relation between $f_1$ and $C_1$ is described as in Equation 4.8,

$$f_1 = maximum\ information\ hiding\ rate \quad \ldots(4.8)$$
$$= mimimum\ information\ hiding\ failure\ rate$$
$$= \frac{C_1}{max\ (C_1, C_2)}$$

The relation between $f_2$ and $C_2$ is explained as following Equation 4.9,

$$f_2 = maximum\ information\ preservation\ rate \quad \ldots(4.9)$$
$$= reciprocal\ of\ minimum\ information\ loss\ rate$$
$$= 1 - \frac{C_2}{max\ (C_1, C_2)}$$

And the relation between $f_3$ and $C_3$ is defined in Equation 4.10,

$$f_3 = maximized\ the\ degree\ of\ non - modification\ rate \quad \ldots(4.10)$$
$$= minimized\ the\ degree\ of\ modification\ rate$$
$$= \frac{C_3}{max\ (C_4)}$$

Now, placing the values of $f_1$, $f_2$, and $f_3$ in Equation 4.7, the final objective function is as in Equation 4.11, adapted from (Mewada et al. 2020) which to be minimized,

$$Min\ f(x) = w_1\left(\frac{C_1}{max[C_1, C_2]}\right) + w_2\left(1 - \frac{C_2}{max[C_1, C_2]}\right) + w_3\left(\frac{C_3}{max(C_4)}\right) \quad \ldots(4.11)$$

Here, the distance amidst individual items set from sanitized and original data is represented by $C_4$. So, the final objective function is derived in Equation 4.12,

$$G = Min\ (f) \quad \ldots(4.12)$$

However, the objective functions $C_1$, $C_2$, $C_3$ are described as minimization of fitness function which is desired in the purpose of medical data preservation. These functions are preferred to determine how efficiently the autism data is sanitized, using the recommended Enhanced Combined PSO-GWO algorithm.

### 4.2.4 The Proposed Enhanced Combined PSO-GWO Framework for Sanitization

In this section, I discussed that how the conventional PSO algorithm as well as GWO algorithm work, respectively. After that I have showed the procedures of working of these algorithms combinedly.

#### a. Working Method of Traditional PSO Algorithm

In the PSO algorithm, there are three vectors. They are x-vector, p-vector, and v-vector. The x-vector keeps track of the present location for the particle in the searching area, wherein the p-vector (pbest) identifies the position of where the particle has discovered the best solution so far. Moreover, the v-vector incorporates particle velocity, indicating where every other particle will move through the following iteration. At the outset, the particles are randomly shifted in specified directions. The particle's orientation might be adjusted gradually, and as a result, it began to move in the direction of the prior best location on its own. After that, it explores the surrounding area for the best locations for some fitness functions, fit = $S^m - S$. Here, the location of the particle is provided as $\vec{M} \in S^m$, while its velocity is provided as $\vec{w}$. Initially, these two variables are picked at random and then updated repeatedly according to two formulae, as shown in Equation 4.13 (W. Li et al. 2021),

$$\vec{w} = \omega\vec{w} + c_1 r_1\left(\vec{q} - \vec{M}\right) + c_2 r_2\left(\vec{f} - \vec{M}\right) \quad \ldots(4.13)$$

In this case, $\omega$, a user-defined behavioural parameter is an inertia weight, which regulates the amount of recurrence in particle velocity. The particle's previous best position (pbest position) is $\vec{q}$, and the particle's previous best position in the swarm (gbest position) is $\vec{f}$, in that way, the particles implicitly interact with each other. And this is weighted using stochastic variables $r_1$, $r_2 \sim U$ (0, 1), while the acceleration constants are $c_1$, $c_2$. Regardless of fitness gains, the velocity is added to the particle's present position to propel it to the next place in the searching area as Equation 4.14,

$$\vec{M} \leftarrow \vec{M} + \vec{w} \quad \ldots(4.14)$$

**b. Working Technique of Traditional GWO Algorithm**

In the GWO algorithm, there has a hierarchical search agent such as level 1(Alpha), level 2 (Beta), level 3(Delta), level 4(Omega). When the grey wolves hunt their prey, then the characteristic of encircling is expressed mathematically as below in Equations 4.15 and 4.16 (Y. Li et al. 2021),

$$\vec{B} = \left| \vec{E}.\vec{M}_q(u) - \vec{M}(u) \right| \quad \ldots(4.15)$$

$$\vec{M}(u+1) = \vec{M}_q(u) - \vec{H}.\vec{B} \quad \ldots(4.16)$$

Here, $u$ is given the current iteration $\vec{H}$ and $\vec{E}$ are referred to as the coefficient vectors. Grey wolves possess a unique skill for detecting the position of the prey and encircle them. These grey wolf hunting actions are mathematically reproduced utilizing alpha, beta, and delta wolves' enhanced awareness of probable prey locations. The first three best solutions are considered, regardless of whether the remainder is required. The mathematical Equations 4.17, 4.18, and 4.19 (Y. Li et al. 2021) are provided as below,

$$\vec{B}_{\alpha} = \left| \vec{E}_1 . \vec{M}_{\alpha} - \vec{M} \right| \quad \ldots(4.17)$$
$$\vec{B}_{\beta} = \left| \vec{E}_2 . \vec{M}_{\beta} - \vec{M} \right|$$
$$\vec{B}_{\delta} = \left| \vec{E}_3 \cdot \vec{M}_{\delta} - \vec{M} \right|$$

$$\vec{M}_1 = \vec{M}_{\alpha} - \vec{H}_1 . (\vec{B}_{\alpha}) \quad \ldots(4.18)$$
$$\vec{M}_2 = \vec{M}_{\beta} - \vec{H}_2 . (\vec{B}_{\beta})$$
$$\vec{M}_3 = \vec{M}_{\delta} - \vec{H}_3 . (\vec{B}_{\delta})$$

$$\vec{M}(u+1) = \frac{\vec{M}_1 + \vec{M}_2 + \vec{M}_3}{3} \quad \ldots(4.19)$$

**c. Working Process of Enhanced Combined PSO-GWO Algorithm**

Despite having a good performance, enhancements can be done on traditional algorithms to address the limitations and improve performance. The traditional PSO algorithm demonstrates a few weaknesses, such as lower performance over a wide range of fields. The GWO algorithm also has a few drawbacks: poorer local searching capability, slower convergence, and lower solving precision. Consequently, further analysis is required to improve robustness and integration.

This study attempts to implement a new hybrid algorithm to solve these issues. The proposed Enhanced Combined PSO-GWO is elaborated as follows: In this regard, the criteria of the PSO algorithm are implemented into the GWO algorithm. The enclosure of the prey mathematical model, in the suggested method, is provided in Equations 4.15 and 4.16, while the mathematical model of the hunting method is shown by the Equations 4.17, 4.18, and 4.19. The updating of the location is the main reformation in the suggested model. So, the updating of the location in this Enhanced Combined PSO-GWO model is shown in Equation 4.20, where $\vec{M}$ refers to the velocity

for the updating of the location of PSO as well as this is demonstrated in Equations 4.13 and 4.14,

$$M(u+1) = \frac{\vec{M}_1 + \vec{M}_2 + \vec{M}_3 + \vec{M}}{4} \quad \ldots(4.20)$$

The optimal key selection based on PSO-GWO is presented in the following Algorithm 4.1.

Algorithm 4.1: Optimal Key Selection by Enhanced Combined PSO-GWO

```
// M_j is the Grey Wolf population where j = 1, 2, ……N. M_α , M_β , and M_δ are the
// best searching agent, 2nd best searching agent, and 3rd best searching agent,
// respectively. Here, e is the components, and H, E are coefficients. The goal of
// this algorithm is to output the best searching agent, M_α .
{
  Set the initial values to the M_j
  Reset e, H, and E
  Measure the fitness values of each searching agent, M_α, M_β, and M_δ.
  while (u < max) do
    {
      for each searching agent, do
        {
          Revise the present location of the searching agent through Equation 4.20.
        }
      Revise e, H, and E
      Assess fitness values for all searching agents
      Revise Mα, M_β, and M_δ.
      u: = u + 1
    }
  return M_α
}
```



## 4.3 SIMULATION AND SYSTEM CONFIGURATION

This section has explained the implementation of the proposed method and their simulation performances. Subsection 4.3.1 mentions the simulation setup, including the autism datasets, and the traditional algorithms that have been compared. The simulation performances compared to those conventional algorithms against various attacks have been demonstrated in subsection 4.3.2.

### 4.3.1 Configuration for Simulation

The proposed method was developed by using the Python programming language. The autism datasets were collected from the faculty of education, Universiti Kebangsaan Malaysia. The autism datasets applied for this study are collected from different aged-group autistic children. These include the autism child dataset at 24 months with 26 attributes and 209 instances, the autism child dataset at 30 months, which have 29 attributes and 209 instances, the autism child dataset at 36 months, including 31 attributes and 234 instances, and the autism child dataset at 48 months, including 33 attributes and 302 instances. All datasets are autism diagnostic data, which have three scoring options, such as z = 0, v = 5, and x = 10. For every type of dataset, the cut-off values were different, at 71, 95, 100, and 105, respectively. The performance of the proposed framework was compared with the existing conventional algorithms such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Crow Search Algorithm (CSA), Differential Evolution (DE), and Adaptive Awareness Probability-based CSA (AAP-CSA).

### 4.3.2 Results and Discussions

There are various types of attacks, such as Known Cipher Attack (KCA), Known Plaintext Attack (KPA), Chosen Cipher Attack (CCA), and Chosen Plaintext Attack (CPA), have been applied on different types of autism datasets. KCA analysis is described to correlate every sanitized data along with its equivalent restored data, whereas KPA is analysed by relating one original data along with the entire original data as well as one sanitized data along with the entire sanitized data. Moreover, CCA attack is stated as an attack model for cryptanalysis wherein the cryptanalyst be able to gather information to attain the decryptions of chosen ciphertexts. The adversary be capable of recovering the hidden secret key used for decryption from that information. Similarly, CPA attack is described as an attack model for cryptanalysis which presumes that the attacker can obtain the ciphertexts for arbitrary plaintexts. The purpose of the attack is obtaining information that decreases the privacy of the encryption scheme. According to definition, the approach that has the lowest value is said to be more secure. Based on these attacks, the simulation has been performed through Figure 4.4 to Figure 4.7.

Initially, the KCA and KPA attacks among the above different sorts of attacks are performed on the proposed framework and other traditional algorithms using two types of autism datasets are simulated in Figure 4.4 to Figure 4.5.

For 24 months autism data, the KCA analysis is investigated at first and shown in Table 4.3.

Table 4.3 KCA analysis on the proposed framework and other existing algorithms while using 24 months autism data

| **Techniques** | **24 months autism dataset** |
|---|---|
| PSO | 0.673110 |
| GA | 0.733333 |
| DE | 0.866666 |
| CSA | 0.225555 |
| AAP-CSA | 0.072323 |
| PSO-GWO | 0.004958 |

From the simulation, the proposed method achieves 99.26%, 99.32%, 99.42%, which are superior to PSO, GA, DE, respectively, whereas 97.80%, and 93.14% greater than CSA as well as AAP-CSA, respectively. The simulation performances are shown in Figure 4.4.

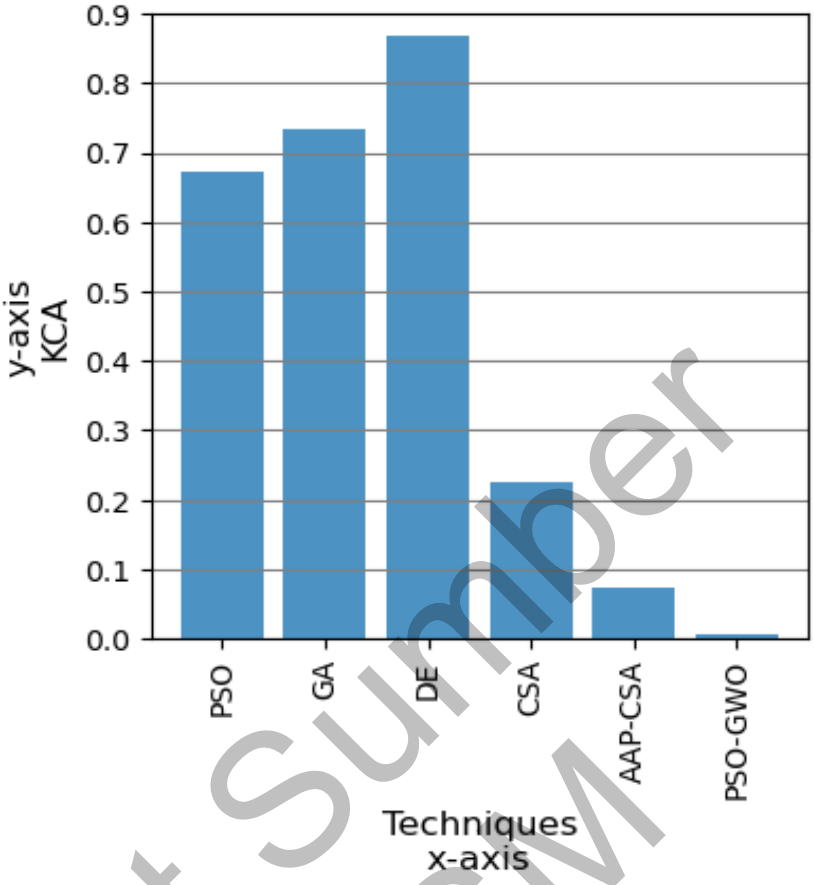



Figure 4.4 Analysis of the performance of various algorithms for the autism at 24 months dataset based on the KCA attack.

Similarly, KPA analysis on the proposed framework and other existing algorithms for 30 months autism data is illustrated in Table 4.4

Table 4.4 KPA analysis on the proposed framework and other existing algorithms while using 30 months autism data

| Techniques | 30 months autism dataset |
|---|---|
| PSO | 0.445000 |
| GA | 0.257553 |
| DE | 0.109694 |
| CSA | 0.079147 |
| AAP-CSA | 0.211810 |
| PSO-GWO | 0.000578 |

For the KPA attack, the proposed method is 99.87%, 99.77%, 99.47%, 99.26% and 99.72% more improved in comparison with the PSO, GA, DE, CSA, and AAP-CSA, respectively which are shown in Figure 4.5.

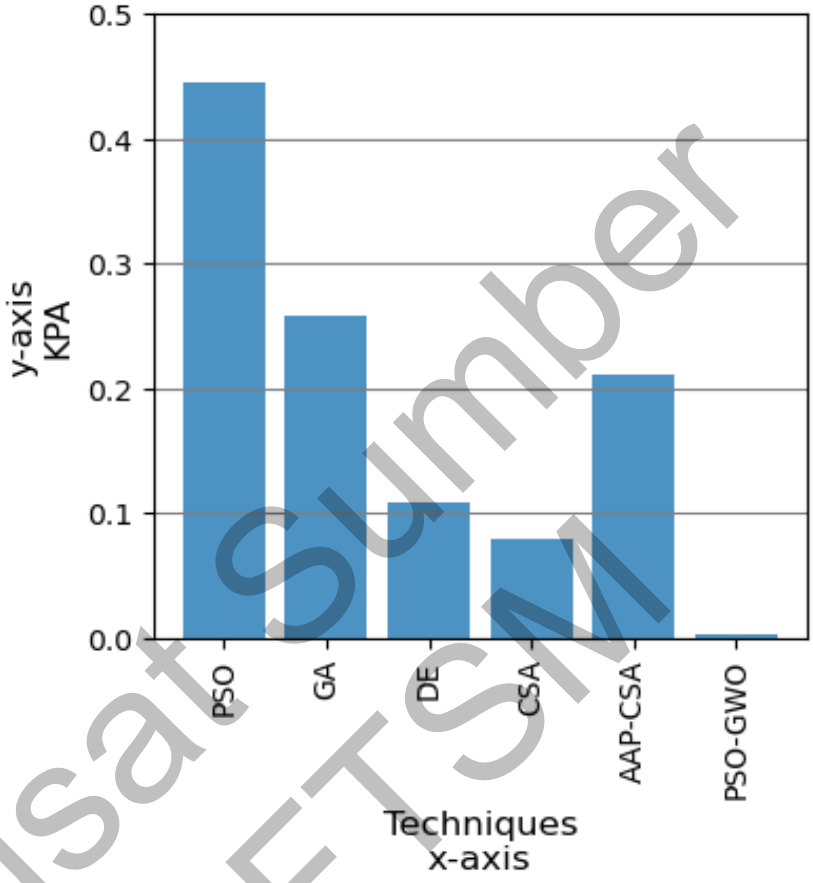



Figure 4.5 Analysis of the performance of various algorithms for the autism at 30 months dataset based on the KPA attack.



The overall performances for KCA and KPA attacks analysis are summarized in Table 4.5.

Table 4.5 The performance of enhanced combined PSO-GWO in terms of KCA and KPA attacks in comparison with the other algorithms under 24- and 30-months autism datasets

| **PSO-GWO** | **PSO** | **GA** | **DE** | **CSA** | **AAP-CSA** | **Attacks** | **Autism datasets** |
|---|---|---|---|---|---|---|---|
| Superior to | 99.26% | 99.32% | 99.42% | 97.80% | 93.14% | KCA | 24 months |
| Higher than | 99.87% | 99.77% | 99.47% | 99.26% | 99.72% | KPA | 30 months |

On the other hand, the CCA and CPA attacks analysis on the proposed framework and other traditional algorithms using another two types of autism datasets are simulated in Figure 4.6 to Figure 4.7.

For 36 months autism data, the CCA analysis on the proposed framework and other existing algorithms is demonstrated in Table 4.6.

Table 4.6 CCA analysis on the proposed framework and other existing algorithms while using 36 months autism data

| **Techniques** | **36 months autism dataset** |
|---|---|
| PSO | 0.08 |
| GA | 0.008989 |
| DE | 0.06 |
| CSA | 0.008998 |
| AAP-CSA | 0.225202 |
| PSO-GWO | 0.007945 |

Figure 4.6 shows that the proposed approach is 90.06%, 11.61%, 86.75%, 11.70% and 96.47% more enhanced from PSO, GA, DE, CSA, and AAP-CSA algorithms, correspondingly.

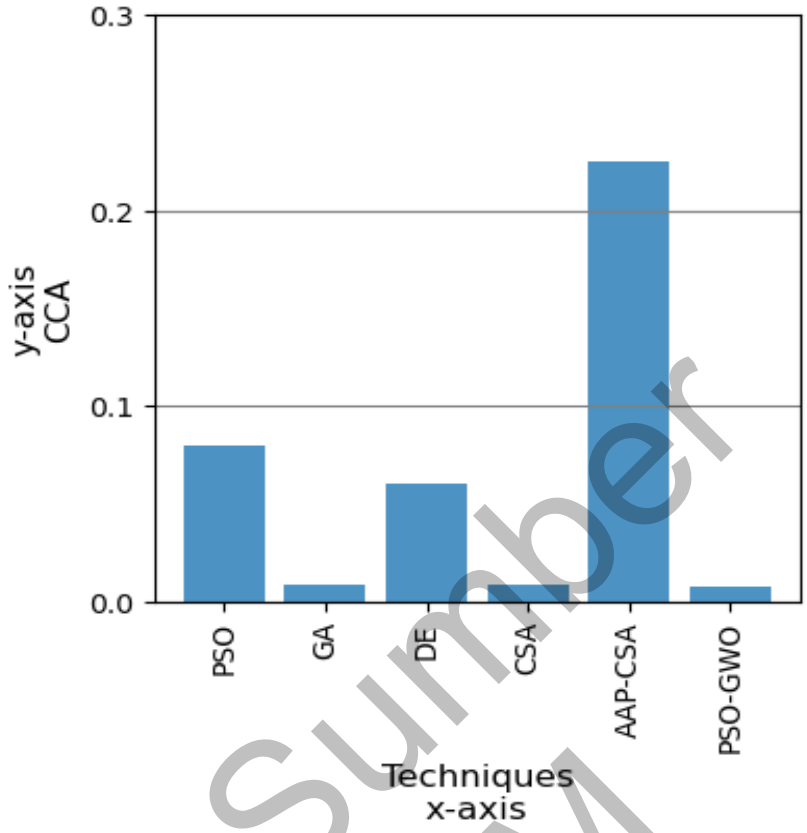



Figure 4.6 Analysis of the performance of various algorithms for the autism at 36 months dataset based on the CCA attack.

Likewise, the CPA analysis on the proposed framework and other existing algorithms for 48 months autism data is exhibited in Table 4.7.

Table 4.7 CPA analysis on the proposed framework and other existing algorithms while using 48 months autism data

| Techniques | 48 months autism dataset |
|---|---|
| PSO | 0.091909 |
| GA | 0.6 |
| DE | 0.2 |
| CSA | 0.3 |
| AAP-CSA | 0.4 |
| PSO-GWO | 0.089876 |

For CPA analysis, this scheme attains 2.21% that is higher than PSO, wherein 85.02%, 55.06%, 70.04% and 77.53% more enhanced than GA, DE, CSA, and AAP-CSA, respectively which is shown in Figure 4.7.

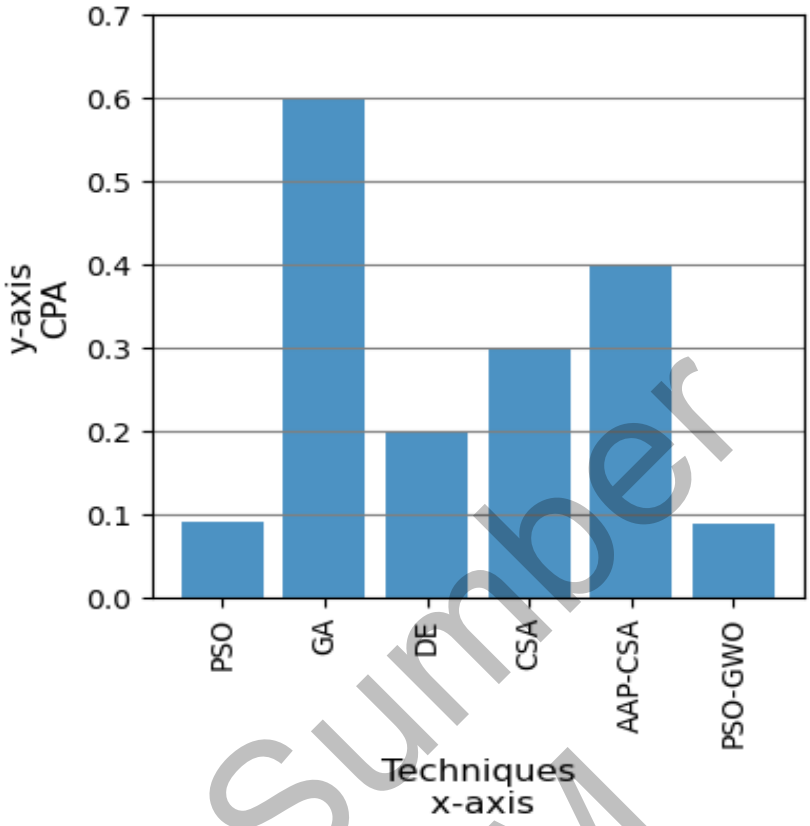



Figure 4.7 Analysis of the performance of various algorithms for the autism at 48 months dataset based on the CPA attack.

The overall outcomes for CCA and CPA attacks analysis are summarized in Table 4.8.

Table 4.8 The performance of enhanced combined PSO-GWO in terms of CCA and CPA attacks in comparison with the other algorithms under the 36- and 48-months autism datasets

| PSO-GWO | PSO | GA | DE | CSA | AAP-CSA | Attacks | Autism datasets |
|---|---|---|---|---|---|---|---|
| Enhanced over | 90.06% | 11.61% | 86.75% | 11.70% | 96.47% | CCA | 36 months |
| Greater than | 2.21% | 85.02% | 55.06% | 70.04% | 77.53% | CPA | 48 months |

## 4.4 DISCUSSIONS

From the above simulation and analysis, the maximum improvement for this technique is 99.26%, 99.32%, 99.42%, which are superior to PSO, GA, DE, respectively, whereas 97.80%, and 93.14% greater than CSA as well as AAP-CSA, respectively for 24 months

autism data by the KCA attack analysis. So, for this kind of attack on 24 months autism datasets, this technique works best over DE which is 99.42%.

In addition, the suggested approach, in the case of KPA attack, attains 99.87%, 99.77%, 99.47%, 99.26% and 99.72% which are more improvement compared to the PSO, GA, DE, CSA, and AAP-CSA, respectively, for the autism at 30 months dataset, wherein the maximum performance is 99.87% over PSO.

For 36 months autism dataset, the simulation result of the proposed technique over CCA attacks achieves 90.06%, 11.61%, 86.75%, 11.70% and 96.47% that are more enhanced from PSO, GA, DE, CSA, and AAP-CSA. So, the best improvement here is over AAP-CSA.

Finally, in terms of the CPA attack under 48 months autism dataset, this technique is 2.21% that is higher than PSO, wherein 85.02%, 55.06%, 70.04% and 77.53% more enhanced than GA, DE, CSA, and AAP-CSA, correspondingly. In this regard, the proposed technique most improved over the GA algorithm.

Thus, the simulation demonstrates that the proposed technique has performed better than the existing conventional algorithms based on some attacks. Therefore, the simulation outcomes reveal that this sanitizing approach performs effectively greater than other traditional algorithms.

Due to the fact that sensitive diagnostic data of autism are critical for determining whether an individual is autistic or not, maintaining privacy of such type of data is critical, which has greater applicability in the healthcare sector. Evidence produced by this study showed that the proposed sanitizing approach conceals these data better than existing algorithms against certain attacks. It is, however, suggested that this recommended approach can be widely applied to the healthcare sector for data privacy.

## 4.5 SUMMARY

The privacy for the autism dataset through sanitizing technique was instigated in this research. The emphasis of this method was to conceal the sensitive data of the patients. Precisely, an optimal key was produced for concealing the sensitive data, which was selected by the proposed Enhanced Combined PSO-GWO framework and employed into the sanitization process to resolve the problems mentioned in introduction. Furthermore, the results obtained by the recommended model were compared with existing traditional algorithms for justification. Mainly, the suggested technique was tested in terms of the different attacks and compared with existing traditional algorithms, and the expected outcomes were achieved, according to the simulation review. This proposed technique, for the 24 months autism dataset in terms of KCA attack, 99.26%, 99.32%, 99.42%, which are superior to PSO, GA, DE, respectively, whereas 97.80%, and 93.14% greater than CSA as well as AAP-CSA, individually. In addition, the suggested approach, in the case of the KPA attack, is 99.87%, 99.77%, 99.47%, 99.26%, and 99.72% more improved over PSO, GA, DE, CSA, and AAP-CSA, respectively, for the autism at 30 months dataset. For 36 months autism dataset, the simulation result of the proposed technique over CCA attack is 90.06%, 11.61%, 86.75%, 11.70% and 96.47% more enhanced from PSO, GA, DE, CSA, and AAP-CSA, correspondingly. Finally, in terms of the CPA attack under 48 months autism dataset, this technique attains 2.21% that is higher than PSO, wherein 85.02%, 55.06%, 70.04% and 77.53% more enhanced than GA, DE, CSA, and AAP-CSA, correspondingly.

Therefore, it is revealed from the analysis that this enhanced technique has an effective performing result over the present conventional algorithms.

# CHAPTER V

# THE DEVELOPMENT OF A RESTORATION PROCESS TO RESTORE A DATABASE FOR AUTISM DATA

## 5.1 INTRODUCTION

The restoration process is a technique where the processed data (archived from the original database) can be retrieved from a sanitized database. Though the different types of privacy preservation models are proposed by many researchers to protect data for the purpose of privacy issues, the optimal key is employed in this restoration process. This optimal key is obtained by the proposed framework. Two optimization algorithms, such as the particle swarm optimization algorithm (PSO) and the grey wolf optimization algorithm (GWO), are applied in this process. The performance of this proposed technique and the existing methods has been measured through some parameters, for instance, information hiding failure rate, information loss rate, and degree of modification rate. After that, by varying the values of the acceleration constants of the optimization algorithms, the performances are also measured through those parameters and more competitive results are achieved in comparison with those existing algorithms. However, the following are the particular contributions made by this research:

- To enhance the data restoration process.
- To improve the restoration performance by utilizing an optimized key by taking into account the aforementioned concerns, which is then utilized in the data sanitization method initially and after that in the restoration process to ensure the privacy and confidentiality of ASD datasets.
- Finally, I compared the performance of the proposed framework to the performance of other existing models and achieved better performances.

## 5.2 ARCHITECTURE FOR RESTORATION OF AUTISM SENSITIVE DATA

Data restoration has been used to preserve the sensitive autism data here. The following Figure 5.1 is the architecture of restoration for autism sensitive data, which protects the confidentiality and privacy of autism data while maintaining the expected results. It is also noted that, from the main overall architecture of the framework in Figure 3.3, the line arrow represents the sanitization process which is one of the focuses of this research, and the dash arrow denotes the restoration process that are another objective of this research. In this section, I have analysed the restoration process and emphasized the privacy issue in respect of autism data. As another significant objective of this research, the restoration of autism sensitive data is illustrated in the following Figure 5.1 and the next subsequent figures in details.

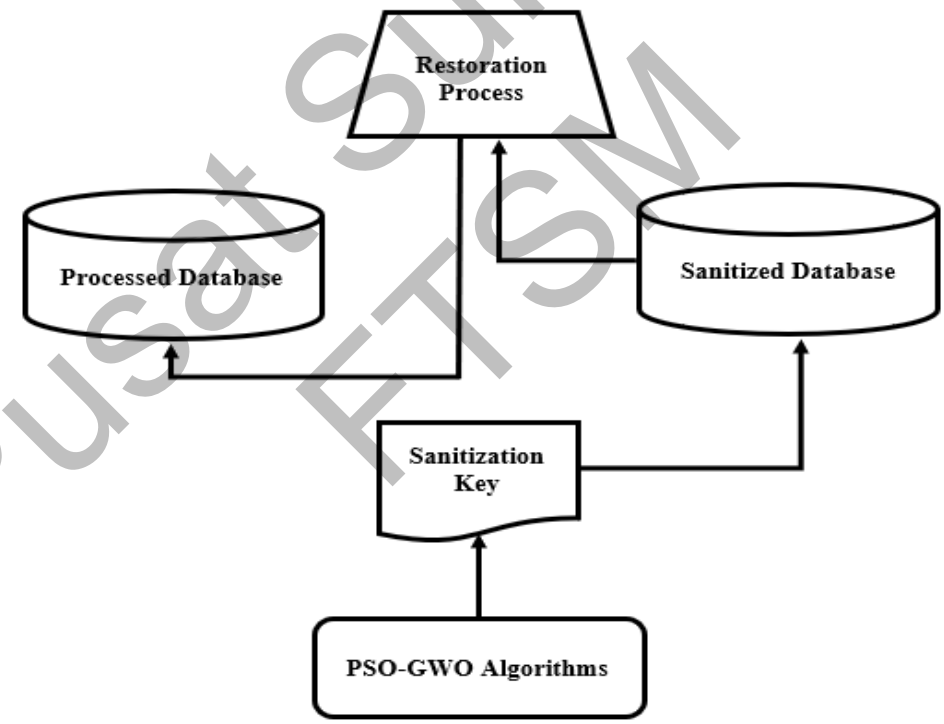



Figure 5.1 Architecture of restoration for autism sensitive data.

This partial architecture from the main overall architecture for restoration process contains the following significant components as below:

1. PSO-GWO Algorithms;
2. Sanitization Key;

3. Sanitized Database;
4. Restoration Process;
5. Processed Database.

Also, the above components are illustrated in the design and development section in chapter III. Furthermore, the mathematical symbols and operators that have been utilized in this process are summarized in the following Table 5.1.

Table 5.1 List of mathematical symbols used in data restoration process

| **Symbols** | **Descriptions** |
|---|---|
| $\widehat{D}$ | Restored Database |
| $D$ | Processed (from original) database |
| $D'$ | Sanitization database |
| $K_1, K_2 \ldots K_N$ | Number of keys |
| $K_2$ | Pruned key matrix |
| $\oplus$ | XOR operator |
| - | Binary Subtraction Operator |
| + | Binary Summation Operator |
| $\otimes$ | Kronecker product |
| $C_1, C_2, C_3$ | Objective functions |
| $f_s$ | Frequency of sensitive itemset in sanitized data |
| $f_m$ | Frequency of sensitive itemset in original data |
| $f_{ns}$ | Frequency of non-sensitive itemset in sanitized data |
| $w_1, w_2, w_3$ | Impact of a particular cost function |
| $f$ | Fitness function |
| $G$ | Minimum objective function |
| $\vec{M}$ | Location of the particle |
| $\vec{w}$ | Velocity of the particle |
| ω | User-defined behavioural parameter (an inertia weight) |
| $\vec{q}$ | Particle's previous best position (pbest position) |
| $\vec{f}$ | Particle's previous best position in the swarm (gbest position) |
| $r_1, r_2$ | Stochastic variables |
| $c_1, c_2$ | Acceleration constants |
| $u$ | Current iteration |
| $\vec{H}, \vec{E}$ | Coefficient vectors |

### 5.2.1 Data Restoration

Figure 5.2 depicts the decoding procedure. From the key generation process and sanitization process, pruned key matrix ($K_2$) and sanitized database ($D´$) are achieved respectively, and is shown in Equation 5.1 (Ahamad et al. 2022),

$$D´ = (K_2 \oplus D) + 1 \quad \ldots(5.1)$$

In this decoding procedure, $D´$ and $K_2$ must be binarized. From binarization block, the sanitized database is minimized by unit input step. In the interim, the XOR operation is performed on that minimized sanitization database and the binarized key matrix, and consequently the restored database is recaptured.

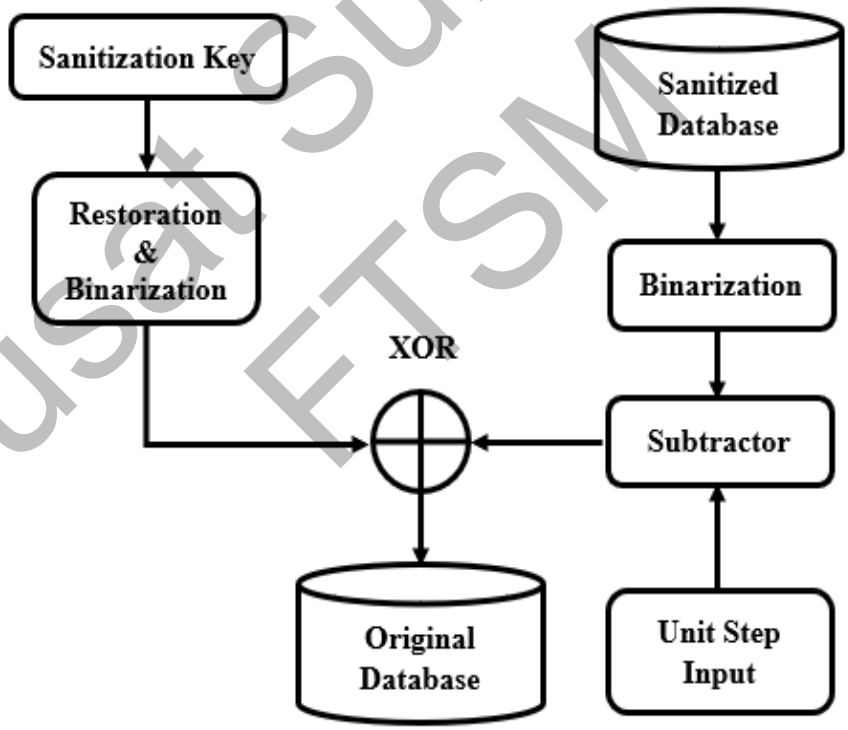



Figure 5.2 Architecture of decoding process

Furthermore, it is noted earlier that the key generation procedure yields sanitized key, which is employed to restore database *D*. This sanitized key is used to generate sanitization database *D´*from where restored database is achieved by use of Equation 5.2 (Ahamad et al. 2022), wherein $\widehat{D}$ implies to restored data.

$$\widehat{D} = (D' - 1) \oplus K_2 \quad \ldots(5.2)$$

And $K_2$ represents the sanitizing key matrix, yielding from $K$. The restoration procedure is described in the following algorithm as Algorithm 5.1.

Algorithm 5.1: Restoration Process by Enhanced Combined PSO-GWO

// $D'$ implies to the sanitizing database, and $K_2$ signifies the sanitizing key while
// $\widehat{D}$ outputs the restored database

**1:** Resultant $K_2$ from transformation.
**2:** Binarization of $K_2$ and $D'$.
**3:** Deduction by unit step input.
**4:** Execution of XOR function.
**5:** Return $\widehat{D}$ .

The performance matrices are information hiding failure rate ($C_1$), information loss rate ($C_2$), and degree of modification ($C_3$), These are also known as the objective functions of this work which are measured through Equation 5.3 to Equation 5.5. In Equation 5.3 (Mewada et al. 2020), $f_s$ and $f_m$ refer to the frequency of sensitive itemset, whereas $f_s$ signifies in the case of sanitized data, and $f_m$ implies in respect of original data.

$$C_1 = \frac{fs}{fm} \quad \ldots(5.3)$$

Likewise, $f_{ns}$ signifies the non-sensitive itemset frequency in reference to sanitized data shown in Equation 5.4 (Mewada et al. 2020),

$$C_2 = \frac{fns}{fm} \quad \ldots(5.4)$$

The Euclidean distance is achieved from Equation 5.5 (Ahamad et al. 2022), while $D$ represents original data, as well as $D'$ denotes sanitized data.

$$C_3 = dist\,(D, D'), \qquad \text{where}, dist \rightarrow Euclidean\ distance \quad \ldots(5.5)$$

Initially, the function is expressed as in Equation 5.6,

$$Max\ f(x) = \sum_{i=1,\ j=1}^{i=M,\ j=N} (w_j\ f_i) \qquad where, \{ f_i : \forall_i = M \} \qquad \{ w_j : \forall_j = N \} \qquad \ldots(5.6)$$

$$= w_1 x_1 + w_2\, x_2 + w_3\, x_3$$

$$= w_1 f_1 + w_2\, f_2 + w_3\, f_3 \qquad where, x = \{ f_1, f_2, f_3 \}$$

Here, x also implies the objective functions, $f_1$, $f_2$, and $f_3$ wherein $f_1$ refers information hiding rate, $f_2$ indicates information preservation rate, and $f_3$ denotes the degree of non-modification rate which all should be maximized. In addition, $w_1$, $w_2$, $w_3$ represent the particular cost function.

Now the relation between $f_1$ and $C_1$ is described as in Equation 5.7,

$$f_1 = maximum\ information\ hiding\ rate \qquad \ldots(5.7)$$
$$= mimimum\ information\ hiding\ failure\ rate$$
$$= \frac{C_1}{max\ (C_1, C_2)}$$

The relation between $f_2$ and $C_2$ is explained as following Equation 5.8,

$$f_2 = maximum\ information\ preservation\ rate \qquad \ldots(5.8)$$
$$= reciprocal\ of\ minimum\ information\ loss\ rate$$
$$= 1 - \frac{C_2}{max\ (C_1, C_2)}$$

And the relation between $f_3$ and $C_3$ is defined in Equation 5.9,

$$f_3 = maximized\ the\ degree\ of\ non - modification\ rate \qquad \ldots(5.9)$$
$$= minimized\ the\ degree\ of\ modification\ rate$$
$$= \frac{C_3}{max\ (C_4)}$$

Now, placing the values of $f_1$, $f_2$, and $f_3$ in Equation 5.6, the final objective function is in Equation 5.10 adapted from (Mewada et al. 2020) which is to be minimized,

$$Min\ f(x) = w_1\left(\frac{C_1}{max[C_1, C_2]}\right) + w_2\left(1 - \frac{C_2}{max[C_1, C_2]}\right) + w_3\left(\frac{C_3}{max(C_4)}\right) \quad \ldots(5.10)$$

Here, $C_4$ denotes the distance amidst individual items set from sanitized and original data. So, the final objective function is derived in Equation 5.11,

$$G = Min\ (f) \quad \ldots(5.11)$$

Specifically, those objective functions $C_1$, $C_2$, and $C_3$ are preferred to determine how efficiently the autism data is restored also by using the recommended Enhanced Combined PSO-GWO framework. So, for data restoration, the objective function of the suggested framework is presented by Equation 5.11.

### 5.2.2 The Proposed Enhanced Combined PSO-GWO Framework for Restoration

For the restoration of autism sensitive data, the working procedures performed by PSO and GWO algorithms combinedly is discussed below:

In the particle swarm optimization, there are three vectors included which are x-vector, p-vector, and v-vector. The x-vector keeps track of the present location for the particle in the searching area, wherein the p-vector (pbest) identifies the position of where the particle has discovered the best solution so far. Moreover, the v-vector incorporates particle velocity, indicating where every other particle will move through the following iteration. At the outset, the particles are randomly shifted in specified directions. The particle's orientation might be adjusted gradually, and as a result, it began to move in the direction of the prior best location on its own. After that, it explores the surrounding area for the best locations for some fitness functions, fit = $S^m - S$. Here, the location of the particle is provided as $\vec{M} \in S^m$, while its velocity is provided as $\vec{w}$.

Initially, these two variables are picked at random and then updated repeatedly according to two formulae, as shown in Equation 5.12 (W. Li et al. 2021),

$$\vec{w} = \omega\vec{w} + c_1 r_1\left(\vec{q} - \vec{M}\right) + c_2 r_2\left(\vec{f} - \vec{M}\right) \quad \ldots(5.12)$$

In this case, $\omega$, a user-defined behavioural parameter is an inertia weight, which regulates the amount of recurrence in particle velocity. The particle's previous best position (pbest position) is $\vec{q}$, and the particle's previous best position in the swarm (gbest position) is $\vec{f}$, in that way, the particles implicitly interact with each other. And this is weighted using stochastic variables $r_1$, $r_2 \sim U(0,1)$, while the acceleration constants are $c_1$, $c_2$. Regardless of fitness gains, the velocity is added to the particle's present position to propel it to the next place in the searching area as Equation 5.13,

$$\vec{M} \leftarrow \vec{M} + \vec{w} \quad \ldots(5.13)$$

Furthermore, in the grey wolf optimization, there have hierarchical search agents such as level 1(Alpha), level 2 (Beta), level 3(Delta), level 4(Omega). When the grey wolves hunt their prey, then the characteristic of encircling is expressed mathematically as below in Equation 5.14 and 5.15 (Y. Li et al. 2021),

$$\vec{B} = \left| \vec{E}.\vec{M}_q(u) - \vec{M}(u) \right| \quad \ldots(5.14)$$

$$\vec{M}(u+1) = \vec{M}_q(u) - \vec{H}.\vec{B} \quad \ldots(5.15)$$

In these equations, $u$ is given the current iteration $\vec{H}$ and $\vec{E}$ are referred to as the coefficient vectors. Grey wolves possess a unique skill for detecting the position of the prey and encircle them. These grey wolf hunting actions are mathematically reproduced utilizing alpha, beta, and delta wolves' enhanced awareness of probable prey locations. The first three best solutions are considered, regardless of whether the remainder is

required. The mathematical Equations 5.16, 5.17, and 5.18 (Y. Li et al. 2021) are provided as below,

$$\vec{B}_\alpha = \left| \vec{E}_1 . \vec{M}_\alpha - \vec{M} \right| \quad \ldots(5.16)$$
$$\vec{B}_\beta = \left| \vec{E}_2 . \vec{M}_\beta - \vec{M} \right|$$
$$\vec{B}_\delta = \left| \vec{E}_3 \cdot \vec{M}_\delta - \vec{M} \right|$$

$$\vec{M}_1 = \vec{M}_\alpha - \vec{H}_1 . (\vec{B}_\alpha) \quad \ldots(5.17)$$
$$\vec{M}_2 = \vec{M}_\beta - \vec{H}_2 . (\vec{B}_\beta)$$
$$\vec{M}_3 = \vec{M}_\delta - \vec{H}_3 . (\vec{B}_\delta)$$

$$\vec{M}(u+1) = \frac{\vec{M}_1 + \vec{M}_2 + \vec{M}_3}{3} \quad \ldots(5.18)$$

In spite of having their performance, enhancements can be done on traditional algorithms to address the limitations and improve performance for restoration of data. The traditional PSO algorithm demonstrates a few weaknesses, such as lower performance over a wide range of fields. The GWO algorithm also has a few drawbacks: poorer local searching capability, slower convergence, and lower solving precision. Consequently, further analysis is required to improve robustness and integration.

This research attempts to implement a new hybrid algorithm to solve this issue. The proposed Enhanced Combined PSO-GWO is elaborated as follows: In this regard, the criteria of the PSO algorithm are implemented into the GWO algorithm. The enclosure of the prey mathematical model, in the suggested method, is provided in Equations 5.14 and 5.15, while the mathematical model of the hunting method is shown by the Equations 5.16, 5.17, and 5.18. The updating of the location is the main reformation in the suggested model. So, the updating of the location in this Enhanced Combined PSO-GWO model is shown in Equation 5.19, where $\vec{M}$ refers to the velocity

for the updating of the location of PSO as well as this is demonstrated in Equations 5.12 and 5.13,

$$M(u+1) = \frac{\vec{M}_1 + \vec{M}_2 + \vec{M}_3 + \vec{M}}{4} \quad \ldots(5.19)$$

More importantly, the $c_1$ and $c_2$ are considered acceleration constants in the traditional PSO algorithm, whereas $c_1$, $c_2$ are fluctuated according to the values 0.1, 0.3, 0.5, 0.7, and 1 in the suggested Enhanced Combined PSO-GWO model. The recovery analysis is assessed based on these varying values and objective functions are evaluated.

## 5.3 SIMULATION AND SYSTEM CONFIGURATION

This section explains how the suggested technique works in terms of restoring data for different types of autism datasets as well as also demonstrates how this technique performed in simulations in varying with the acceleration constants, $c_1$ and $c_2$.

### 5.3.1 Simulation Setup

The recommended method for restoration was experimented by utilizing the Python programming language. For the experiments in order to restoration, the datasets of autism were also accumulated from the faculty of education, Universiti Kebangsaan Malaysia. The autism datasets employed for this work are also different aged autistic children, for instance, autism child dataset 24 months, autism child dataset 30 months, autism child dataset 36 months, autism child dataset 48 months with various attributes and instances. Finally, the output from this experiment were compared with the conventional algorithms, for example, Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Crow Search Algorithm (CSA), Differential Evolution (DE), and Adaptive Awareness Probability-based CSA (AAP-CSA) as well and able to acquire competitive results.

### 5.3.2 Results and Discussions

This subsection focuses the restoration analysis of autism data by the proposed framework and compares with the existing models first. After that this framework also shows the impact of restoration on varying values of acceleration constants, $c_1$ and $c_2$.

#### a. Restoration Analysis by the Proposed Framework over Conventional Algorithms

Table 5.2, Table 5.3, Table 5.4, and Table 5.5 demonstrate the performance analysis by restoration procedure of the Enhanced Combined PSO-GWO framework for the four autism datasets.

Initially, for 24 months autism child dataset, the suggested framework in case of $C_1$ performs 99.26%, 99.32%, 99.43%, 97.80%, and 93.14% greater than PSO, GA, DE, CSA, as well as AAP-CSA, correspondingly and is shown in Table 5.2. This model also depicts the performance over GA by 72.24% for $C_3$. And also, the enhanced model performs 43.60% better than AAP-CSA for $f$.

Table 5.2 Analysis on recovery for 24 months autism child dataset

| Functions | PSO | GA | DE | CSA | AAP-CSA | PSO-GWO |
|---|---|---|---|---|---|---|
| $C_1$ | 6.73110417 | 7.333332 | 8.666669 | 2.255556 | 0.723232 | 0.049582688 |
| $C_2$ | 0.96665588 | 0.985866 | 0.899999 | 0.899989 | 1.002099 | 1.007415858 |
| $C_3$ | 450.040250 | 2199.030 | 470.5105 | 400.0021 | 490.3019 | 610.5032997 |
| $f$ | 31.1909070 | 3.906688 | 19.50060 | 0.713161 | 60.10929 | 33.89987687 |

The performance analysis of the recommended framework for 30 months autism child dataset over the existing algorithms is shown in Table 5.3. The framework reveals that 99.87%, 99.78%, 99.47%, 99.27% and 99.73% enhanced over PSO, GA, DE, CSA, and AAP-CSA, correspondingly for $C_1$. The model shows 48.56% better than CSA for $C_2$ whereas 41% and 9.66% improved over GA and CSA, respectively for $C_3$.

Table 5.3 Analysis on recovery for 30 months autism child dataset

| Functions | PSO | GA | DE | CSA | AAP-CSA | PSO-GWO |
|---|---|---|---|---|---|---|
| $C_1$ | 4.45 | 2.57553 | 1.096949 | 0.791470 | 2.11810 | 0.005785754 |
| $C_2$ | 0.8922342 | 0.999245 | 0.898999 | 2.005758 | 0.899899 | 1.031786858 |
| $C_3$ | 1210.2510 | 2390.050 | 1310.297 | 1560.998 | 1055.604 | 1410.203456 |
| $f$ | 31.2160208 | 22.43441 | 59.50199 | 4.521818 | 49.60847 | 65.88978746 |

Table 5.4 illustrates the improvement results of the recommended model by applying 36 months autism child dataset with regard to $C_1$ which is 90.07%, 11.62%, 86.76%, 11.70% and 96.47% excellent over PSO, GA, DE, CSA, and AAP-CSA, individually. Again, this model shows 48.98%, 51.39%, 51.67%, 51.74% more enhancement over PSO, GA, DE, CSA, respectively for $C_2$ and 8.74% over GA for $C_3$. After that the model also shows 29.74%, 73.68%, and 41.53% better from PSO, DE and AAP-CSA, respectively, in case of $f$.

Table 5.4 Analysis on recovery for 36 months autism child dataset

| Functions | PSO | GA | DE | CSA | AAP-CSA | PSO-GWO |
|---|---|---|---|---|---|---|
| $C_1$ | 0.80 | 0.089899 | 0.6 | 0.089989 | 2.252020 | 0.079456492 |
| $C_2$ | 2.00092711 | 2.100122 | 2.112323 | 2.115234 | 0.889812 | 1.020892055 |
| $C_3$ | 6299.98897 | 9040.232 | 8001.009 | 5999.876 | 4499.201 | 8250.055682 |
| $f$ | 150.080677 | 17.42103 | 400.7061 | 40.06892 | 180.3530 | 105.4525589 |

Finally, Table 5.5 reveals for the 48 months autism child dataset that the model performs 2.21%, 85.02%, 55.06%, 70.04% and 77.53% higher than PSO, GA, DE, CSA, and AAP-CSA respectively in respect of $C_1$. This framework also shows 45.30%, 47.57%, 45.48%, 45.80% and 45.48% greater than PSO, GA, DE, CSA, and AAP-CSA respectively for $C_2$. In terms of $C_3$, the model outperforms in comparison with the PSO, DE and AAP-CSA by 31.55%, 3.32%, and 7.55%, correspondingly. This recommended framework also achieves 23.19%, 66.94%, and 71.03% superior to PSO, DE and AAP-CSA, respectively, in case of $f$.

Table 5.5 Analysis on recovery for 48 months autism child dataset

| Functions | PSO | GA | DE | CSA | AAP-CSA | PSO-GWO |
|---|---|---|---|---|---|---|
| $C_1$ | 0.0919090 | 0.6 | 0.2 | 0.3 | 0.4 | 0.089876584 |
| $C_2$ | 2.0033536 | 2.090031 | 2.009987 | 2.022078 | 2.009948 | 1.095866585 |
| $C_3$ | 13120.440 | 4010.120 | 9289.420 | 8570.42 | 9714.452 | 8980.896565 |
| $f$ | 150.896020 | 20.39320 | 350.6099 | 10.799127 | 400.1057 | 115.9087576 |

As a consequence, the restoration process of the suggested model outperforms other traditional algorithms, as demonstrated by the results above.

**b. Impact for Different $c_1$ and $c_2$ Values**

The restoration of autism data had been measured based on objective functions of the Enhanced Combined PSO-GWO framework. Here, the acceleration constants had been updated by varying as, $0 < (c_1, c_2) <= 1$ and are obtained by the Equation 5.12. Performances of analysis on cost functions for four types of autism child datasets based on these varying values have been revealed in Figure 5.3 to Figure 5.22.

At first, by taking the 24 months autism child dataset with the values for $c_1$ = 0.1 and $c_2$ = 0.1, the outcomes of GA for objective functions, $C_1$, $C_2$, $C_3$ and $f$ are 7.333, 0.985, 2199.030 and 3.906, while the proposed technique achieves 0.049, 1.007, 590.40, and 32.80 correspondingly, and shown in Table 5.6.

Table 5.6 Cost analysis for 24 months autism data, while $c_1$ = 0.1 and $c_2$ = 0.1

| Objective Functions | GA | PSO-GWO |
|---|---|---|
| $C_1$ | 7.333 | 0.049 |
| $C_2$ | 0.985 | 1.007 |
| $C_3$ | 2199.030 | 590.40 |
| $f$ | 3.906 | 32.80 |

The simulation is shown in Figure 5.3, where the proposed technique is 99.33%, 73.15% more improved to GA for $C_1$, and $C_3$.

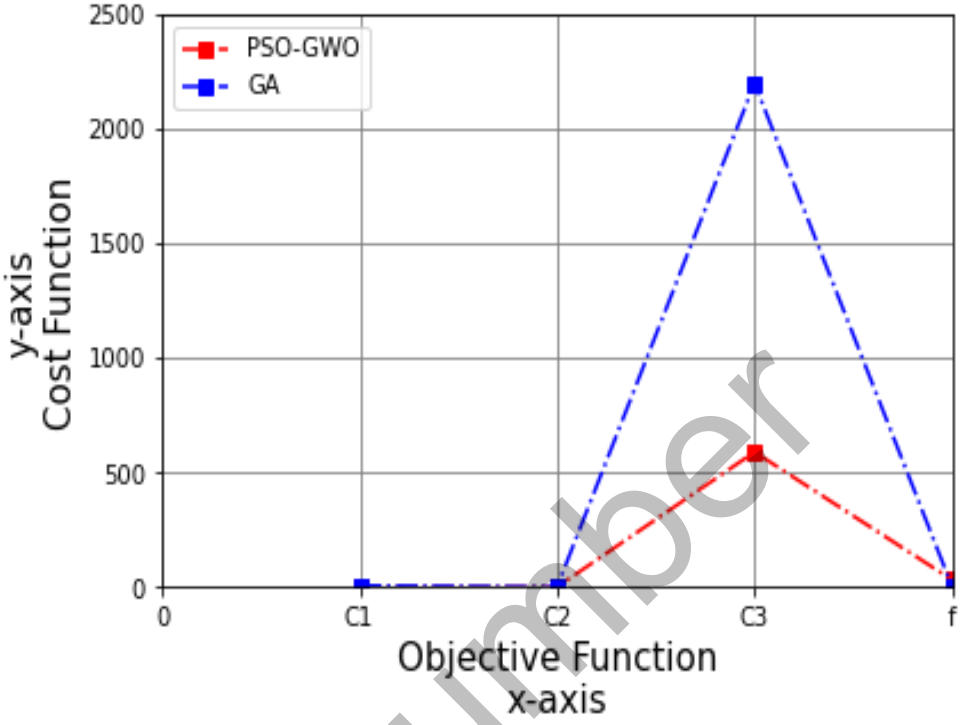



Figure 5.3 Performance analysis on cost function for autism child dataset 24 months, while $c_1 = 0.1$ and $c_2 = 0.1$.

Similarly, the values of $C_1$, $C_2$, $C_3$ and $f$ for GA are 2.575, 0.999, 2390.050 and 22.434, whereas CSA are 0.791, 2.005, 1560.998 and 4.521, and the proposed method are 0.005, 1.031, 1340.10 and 55.20, respectively for 30 months autism child dataset when the values of acceleration constants are $c_1 = 0.1$ and $c_2 = 0.1$ and summarized in Table 5.7.

Table 5.7 Cost analysis for 30 months autism data, while $c_1 = 0.1$ and $c_2 = 0.1$

| Objective Functions | GA | CSA | PSO-GWO |
|---|---|---|---|
| $C_1$ | 2.575 | 0.791 | 0.005 |
| $C_2$ | 0.999 | 2.005 | 1.031 |
| $C_3$ | 2390.050 | 1560.998 | 1340.10 |
| $f$ | 22.434 | 4.521 | 55.20 |

So, the suggested technique shows 99.81%, and 43.93% higher than GA for $C_1$, and $C_3$ correspondingly, whereas 99.37%, 48.58%, and 14.15% greater than CSA for $C_1$, $C_2$, and $C_3$, respectively, which are simulated in Figure 5.4.

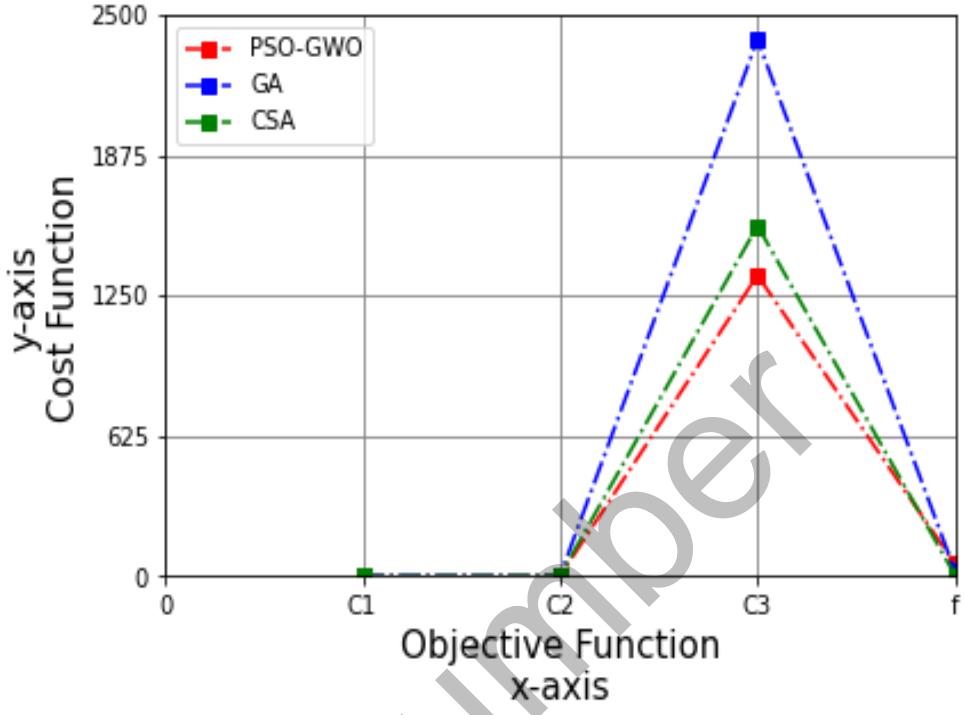



Figure 5.4 Performance analysis on cost function for autism child dataset 30 months, while $c_1 = 0.1$ and $c_2 = 0.1$.

For 36 months autism child dataset with the values, $c_1 = 0.1$ and $c_2 = 0.1$, the objective functions, $C_1$, $C_2$, $C_3$, and $f$ for the GA are 0.089, 2.100, 9040.232 and 17.421 wherein the proposed scheme assessed 0.079, 1.020, 8140.45 and 100.30 correspondingly, that are revealed in Table 5.8.

Table 5.8 Cost analysis for 36 months autism data, while $c_1 = 0.1$ and $c_2 = 0.1$

| **Objective Functions** | **GA** | **PSO-GWO** |
|---|---|---|
| $C_1$ | 0.089 | 0.079 |
| $C_2$ | 2.100 | 1.020 |
| $C_3$ | 9040.232 | 8140.45 |
| $f$ | 17.421 | 100.30 |

Figure 5.5 illustrated that the proposed method is 11.24%, 51.43%, 9.95% more enhanced over GA for $C_1$, $C_2$, and $C_3$.

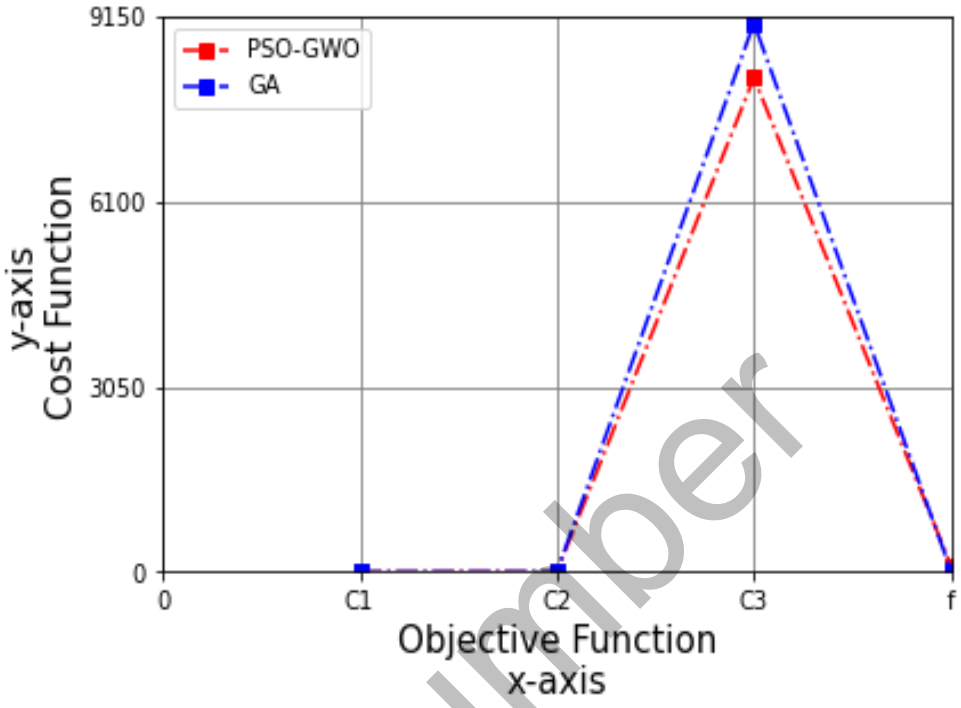



Figure 5.5 Performance analysis on cost function for autism child dataset 36 months, while $c_1 = 0.1$ and $c_2 = 0.1$.

In case of 48 months autism child dataset regarding the acceleration constants, $c_1 = 0.1$ and $c_2 = 0.1$, the $C_1$, $C_2$, $C_3$ and $f$ of PSO are 0.091, 2.003, 13020.30 and 140.60, respectively; DE are 0.2, 2.009, 9289.420 and 350.609, and AAP-CSA are 0.4, 2.009, 9714.452, and 400.105; whereas the suggested framework attains 0.089, 1.095, 8750.40 and 105.80, respectively, which are shown in Table 5.9.

Table 5.9 Cost analysis for 48 months autism data, while $c_1 = 0.1$ and $c_2 = 0.1$

| Objective Functions | PSO | DE | AAP-CSA | PSO-GWO |
|---|---|---|---|---|
| $C_1$ | 0.091 | 0.2 | 0.4 | 0.089 |
| $C_2$ | 2.003 | 2.009 | 2.009 | 1.095 |
| $C_3$ | 13020.30 | 9289.420 | 9714.452 | 8750.40 |
| $f$ | 140.60 | 350.609 | 400.105 | 105.80 |

Figure 5.6 shows that the PSO-GWO is higher than PSO by 2.20%, 45.33%, 32.79% and 24.75%, DE by 55.5%, 45.50%, 5.80% and 69.82%, and AAP-CSA by 77.75%, 45.50%, 9.92%, and 73.56% for $C_1$, $C_2$, $C_3$ and $f$.

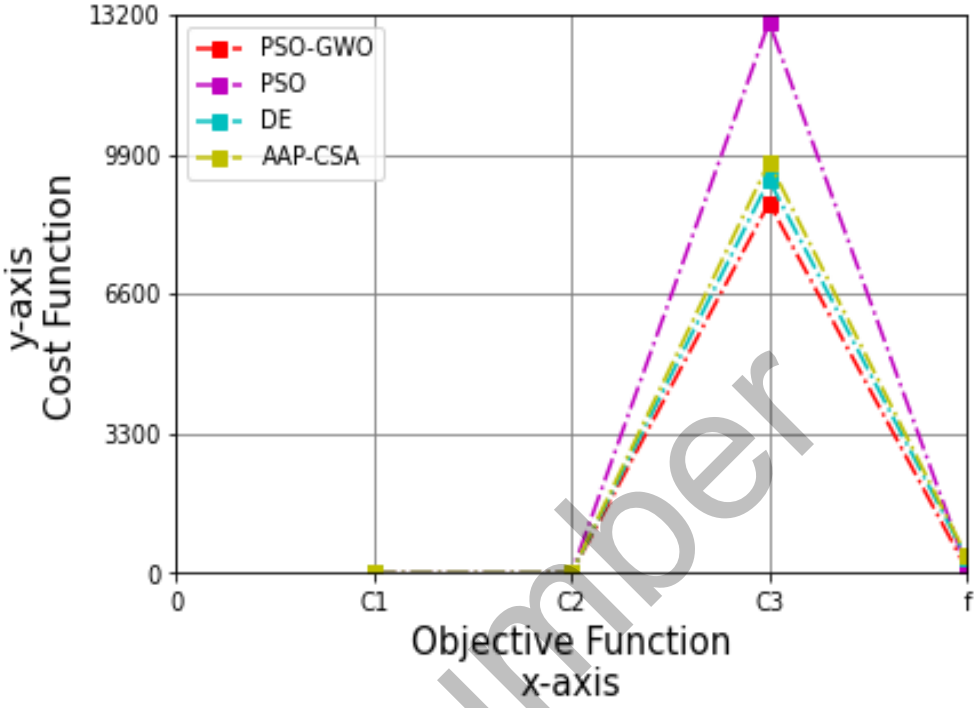



Figure 5.6 Performance analysis on cost function for autism child dataset 48 months, while $c_1 = 0.1$ and $c_2 = 0.1$.

After that, the values of the acceleration constants are set to $c_1 = 0.3$ and $c_2 = 0.3$, so the objective functions, $C_1$, $C_2$, $C_3$ and $f$ for GA are 7.333, 0.985, 2199.030 and 3.906, while the proposed technique achieves 0.049, 1.007, 500.25, and 30.05, respectively, over the 24 months autism child dataset and is shown in Table 5.10.

Table 5.10 Cost analysis for 24 months autism data, while $c_1 = 0.3$ and $c_2 = 0.3$

| **Objective Functions** | **GA** | **PSO-GWO** |
|---|---|---|
| $C_1$ | 7.333 | 0.049 |
| $C_2$ | 0.985 | 1.007 |
| $C_3$ | 2199.030 | 500.25 |
| $f$ | 3.906 | 30.05 |

The simulation is presented in Figure 5.7, where the proposed technique is 99.33%, 77.25% more improved to GA for $C_1$, and $C_3$.

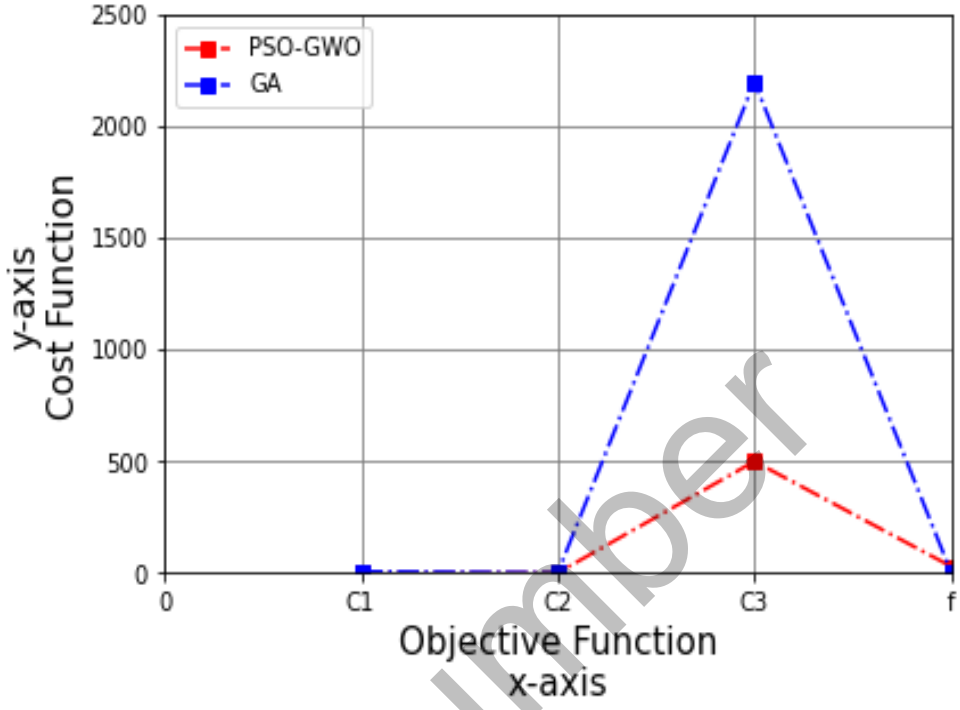



Figure 5.7 Performance analysis on cost function for autism child dataset 24 months, while $c_1 = 0.3$ and $c_2 = 0.3$.

Under the 30 months autism child dataset with respect to $c_1 = 0.3$ and $c_2 = 0.3$, the values of $C_1$, $C_2$, $C_3$ and $f$ for GA are 2.575, 0.999, 2390.050 and 22.434, DE are 1.096, 0.898, 1310.297, and 59.501, CSA are 0.791, 2.005, 1560.998 and 4.521, whereas the proposed method attains 0.005, 1.031, 1240.10 and 45.15, respectively, that are brief in Table 5.11.

Table 5.11 Cost analysis for 30 months autism data, while $c_1 = 0.3$ and $c_2 = 0.3$

| Objective Functions | GA | DE | CSA | PSO-GWO |
|---|---|---|---|---|
| $C_1$ | 2.575 | 1.096 | 0.791 | 0.005 |
| $C_2$ | 0.999 | 0.898 | 2.005 | 1.031 |
| $C_3$ | 2390.050 | 1310.297 | 1560.998 | 1240.10 |
| $f$ | 22.434 | 59.501 | 4.521 | 45.15 |

The simulation is revealed in Figure 5.8. Here, the suggested method is 99.81%, and 48.11% superior to GA for $C_1$ and $C_3$, and 99.54%, 5.36%, and 24.12% better than DE for $C_1$, $C_3$, and $f$, as well as 99.37%, 48.58%, and 20.56% greater than CSA for $C_1$, $C_2$, and $C_3$.

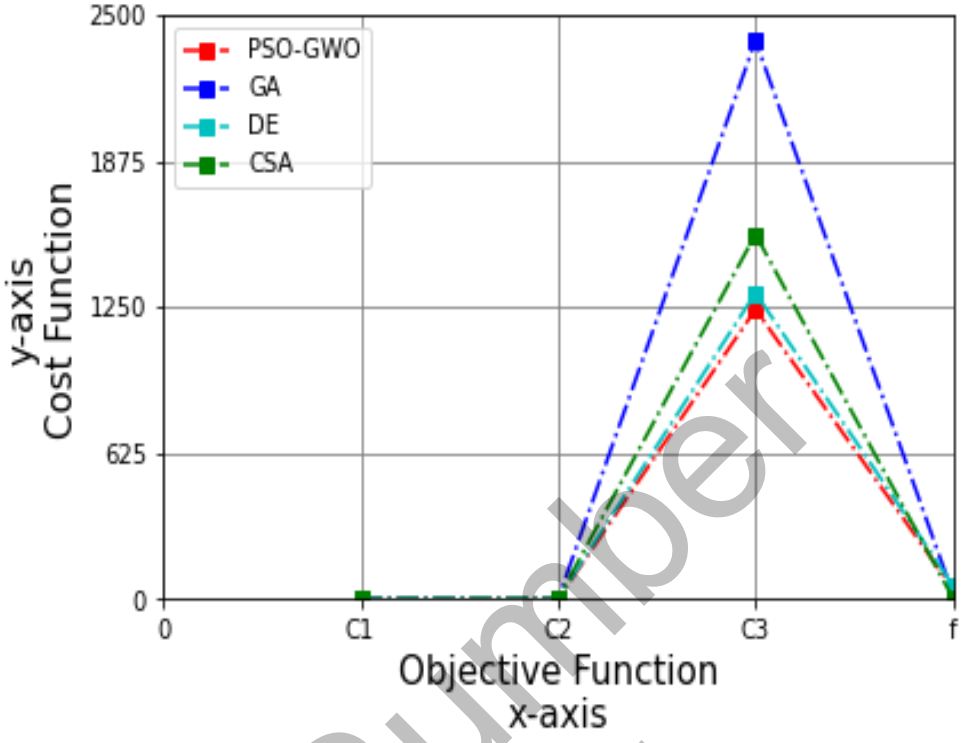



Figure 5.8 Performance analysis on cost function for autism child dataset 30 months, while $c_1 = 0.3$ and $c_2 = 0.3$.

For 36 months autism child dataset in terms of $c_1 = 0.3$ and $c_2 = 0.3$, the objective functions, $C_1$, $C_2$, $C_3$, and $f$ for the GA are 0.089, 2.100, 9040.232 and 17.421, DE are 0.6, 2.112, 8001.009, and 400.706, wherein the attaining outcomes are 0.079, 1.020, 7900.25 and 90.20 correspondingly, that are presented in Table 5.12.

Table 5.12 Cost analysis for 36 months autism data, while $c_1 = 0.3$ and $c_2 = 0.3$

| Objective Functions | GA | DE | PSO-GWO |
|---|---|---|---|
| $C_1$ | 0.089 | 0.6 | 0.079 |
| $C_2$ | 2.100 | 2.112 | 1.020 |
| $C_3$ | 9040.232 | 8001.009 | 7900.25 |
| $f$ | 17.421 | 400.706 | 90.20 |

The recommended method is 11.24%, 51.43%, and 12.61% higher than GA for $C_1$, $C_2$, and $C_3$, wherein 86.83%, 51.70%, 1.26%, and 77.49% greater than DE for $C_1$, $C_2$, $C_3$, and $f$, respectively, which are demonstrated in Figure 5.9.

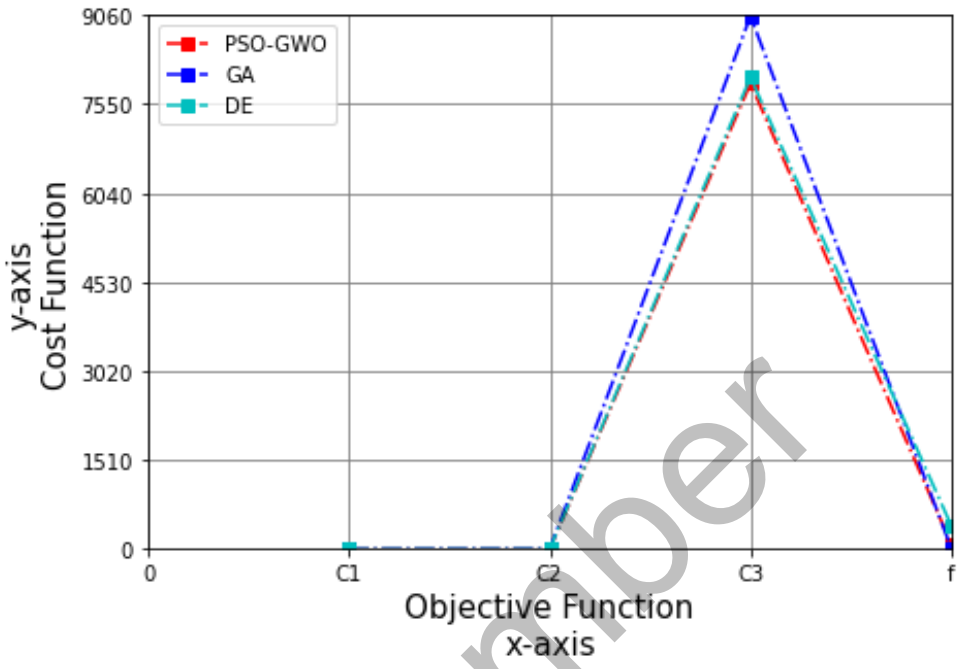



Figure 5.9 Performance analysis on cost function for autism child dataset 36 months, while $c_1 = 0.3$ and $c_2 = 0.3$.

Setting the acceleration constants values, $c_1 = 0.3$ and $c_2 = 0.3$, under the 48 months autism child dataset, the results of $C_1$, $C_2$, $C_3$ and $f$ for PSO are 0.091, 2.003, 12340.20 and 100.40, respectively; DE are 0.2, 2.009, 9289.420 and 350.609; CSA are 0.3, 2.022, 8570.42, and 10.799, and AAP-CSA are 0.4, 2.009, 9714.452, and 400.105; whereas the suggested framework achieves 0.089, 1.095, 8240.20 and 85.60, respectively, which are revealed in Table 5.13.

Table 5.13 Cost analysis for 48 months autism data, while $c_1 = 0.3$ and $c_2 = 0.3$

| Objective Functions | PSO | DE | CSA | AAP-CSA | PSO-GWO |
|---|---|---|---|---|---|
| $C_1$ | 0.091 | 0.2 | 0.3 | 0.4 | 0.089 |
| $C_2$ | 2.003 | 2.009 | 2.022 | 2.009 | 1.095 |
| $C_3$ | 12340.20 | 9289.420 | 8570.42 | 9714.452 | 8240.20 |
| $f$ | 100.40 | 350.609 | 10.799 | 400.105 | 85.60 |

The simulation is illustrated in Figure 5.10, and it is revealed that this technique 2.20%, 45.33%, 33.22%, and 14.74% higher than PSO, 55.5%, 45.50%, 11.29%, and 75.59% greater than DE, 77.75%, 45.50%, 15.18%, and 78.61% superior to AAP-CSA for $C_1$, $C_2$, $C_3$ and $f$, whereas 70.33%, 45.85%, and 3.85% better than CSA for $C_1$, $C_2$, and $C_3$, respectively.

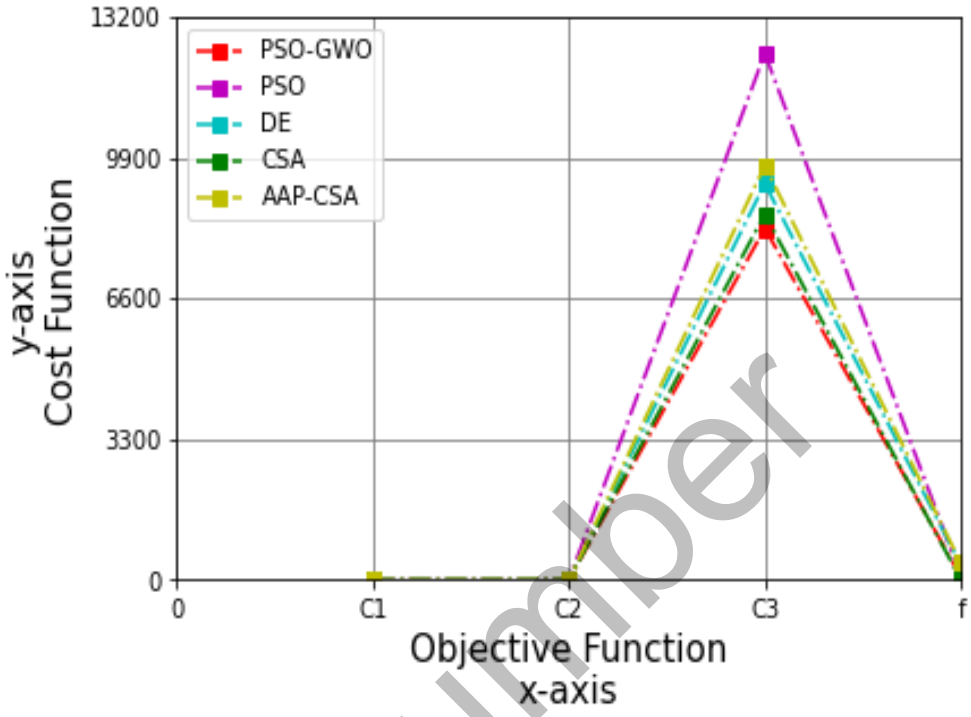



Figure 5.10 Performance analysis on cost function for autism child dataset 48 months, while $c_1 = 0.3$ and $c_2 = 0.3$.

Correspondingly, for the 24 months autism child dataset in case of $c_1 = 0.5$ and $c_2 = 0.5$, the outcomes of $C_1$, $C_2$, $C_3$ and $f$ for PSO are 6.731, 0.966, 440.10, and 24.15, GA are 7.333, 0.985, 2199.030 and 3.906, DE are 8.666, 0.899, 470.51, and 19.5, AAP-CSA are 0.723, 1.002, 490.301, and 60.109, while the proposed technique achieves 0.049, 1.007, 401.15, and 25.10, respectively, that is shown in Table 5.14.

Table 5.14 Cost analysis for 24 months autism data, while $c_1 = 0.5$ and $c_2 = 0.5$

| Objective Functions | PSO | GA | DE | AAP-CSA | PSO-GWO |
|---|---|---|---|---|---|
| $C_1$ | 6.731 | 7.333 | 8.666 | 0.723 | 0.049 |
| $C_2$ | 0.966 | 0.985 | 0.899 | 1.002 | 1.007 |
| $C_3$ | 440.10 | 2199.030 | 470.51 | 490.301 | 401.15 |
| $f$ | 24.15 | 3.906 | 19.5 | 60.109 | 25.10 |

Figure 5.11 demonstrates that the suggested method is 99.27%, 8.85% better than PSO, 99.33%, 81.76% higher than GA, 99.43%, 14.74% superior to DE for $C_1$, and $C_3$, whereas 93.22%, 18.18%, and 58.24% greater than AAP-CSA for $C_1$, $C_3$ and $f$, respectively.

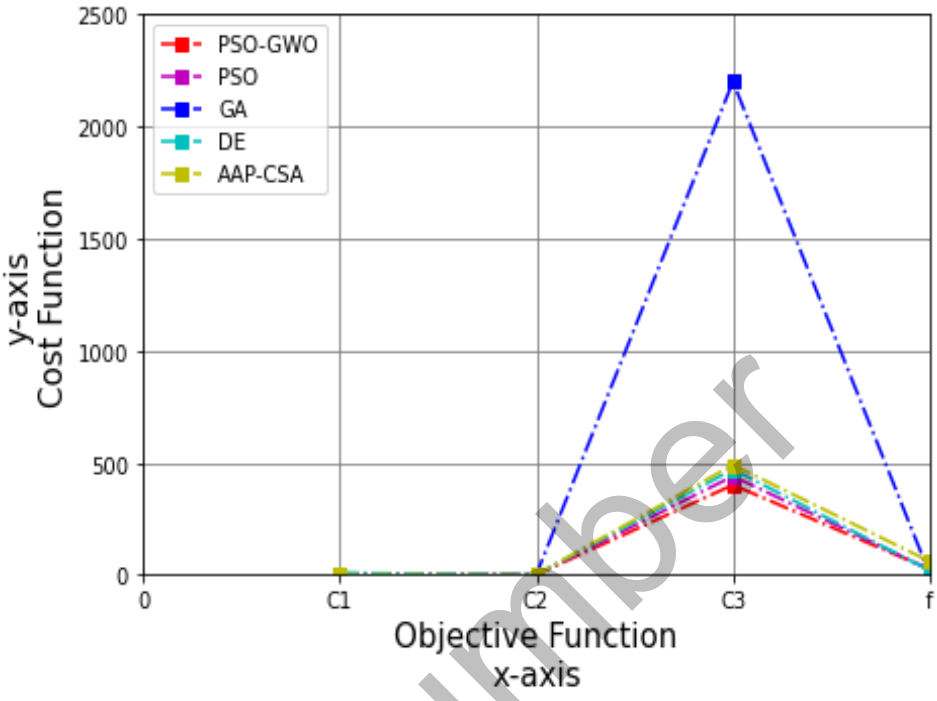



Figure 5.11 Performance analysis on cost function for autism child dataset 24 months, while $c_1 = 0.5$ and $c_2 = 0.5$.

Moreover, for the 30 months autism child dataset with regarding to $c_1 = 0.5$ and $c_2 = 0.5$, the values of $C_1$, $C_2$, $C_3$ and $f$ for PSO are 4.45, 0.892, 1150.15, and 25.11, GA are 2.575, 0.999, 2390.050 and 22.434, DE are 1.096, 0.898, 1310.297, and 59.501, CSA are 0.791, 2.005, 1560.998 and 4.521, whereas the proposed method conquers 0.005, 1.031, 1100.20 and 30.25, respectively, that are brief in Table 5.15.

Table 5.15 Cost analysis for 30 months autism data, while $c_1 = 0.5$ and $c_2 = 0.5$

| Objective Functions | PSO | GA | DE | CSA | PSO-GWO |
|---|---|---|---|---|---|
| $C_1$ | 4.45 | 2.575 | 1.096 | 0.791 | 0.005 |
| $C_2$ | 0.892 | 0.999 | 0.898 | 2.005 | 1.031 |
| $C_3$ | 1150.15 | 2390.050 | 1310.297 | 1560.998 | 1100.20 |
| $f$ | 25.11 | 22.434 | 59.501 | 4.521 | 30.25 |

The simulation is revealed in Figure 5.12. Here, the recommended technique is 99.89%, 4.34% greater than PSO, 99.81%, and 53.97% superior to GA for $C_1$ and $C_3$, and 99.54%, 16.03%, and 49.16% better than DE for $C_1$, $C_3$, and $f$, as well as 99.37%, 48.58%, and 29.52% greater than CSA for $C_1$, $C_2$, and $C_3$, respectively.

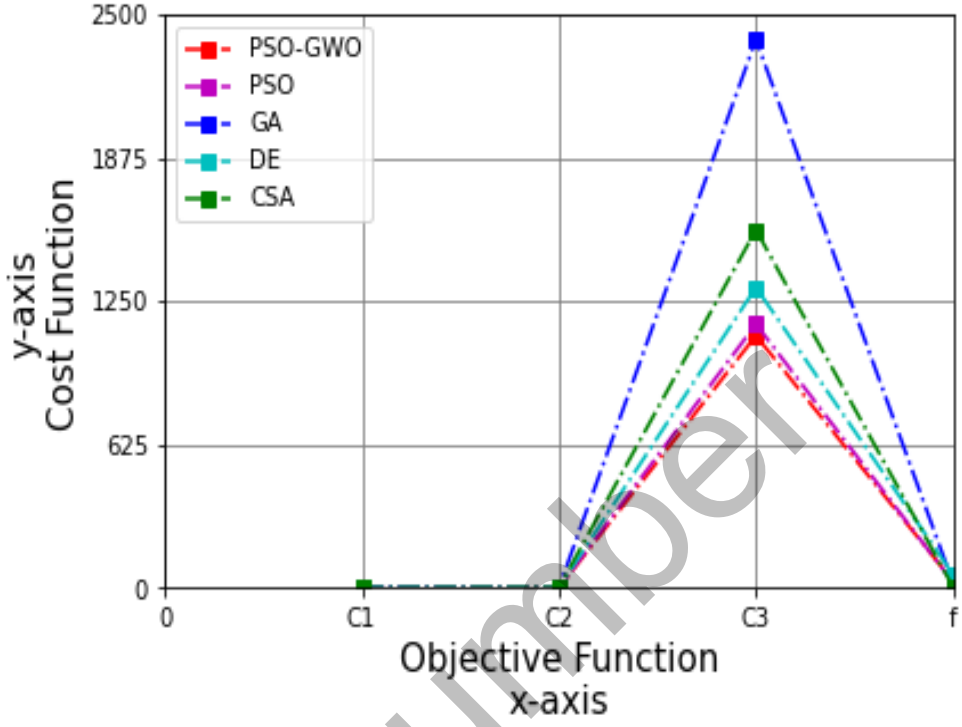



Figure 5.12 Performance analysis on cost function for autism child dataset 30 months, while $c_1 = 0.5$ and $c_2 = 0.5$.

Applying the 36 months autism child dataset in respect of acceleration constant values $c_1 = 0.5$ and $c_2 = 0.5$, the objective functions, $C_1$, $C_2$, $C_3$, and $f$ for the GA are 0.089, 2.100, 9040.232 and 17.421, DE are 0.6, 2.112, 8001.009, and 400.706, wherein the attaining outcomes are 0.079, 1.020, 7100.45 and 65.10 correspondingly, that are presented in Table 5.16.

Table 5.16 Cost analysis for 36 months autism data, while $c_1 = 0.5$ and $c_2 = 0.5$

| Objective Functions | GA | DE | PSO-GWO |
|---|---|---|---|
| $C_1$ | 0.089 | 0.6 | 0.079 |
| $C_2$ | 2.100 | 2.112 | 1.020 |
| $C_3$ | 9040.232 | 8001.009 | 7100.45 |
| $f$ | 17.421 | 400.706 | 65.10 |

The suggested scheme is 11.24%, 51.43%, and 21.46% higher than GA for $C_1$, $C_2$, and $C_3$, wherein 86.83%, 51.70%, 11.26%, and 83.75% greater than DE for $C_1$, $C_2$, $C_3$, and $f$, respectively, which are demonstrated in Figure 5.13.

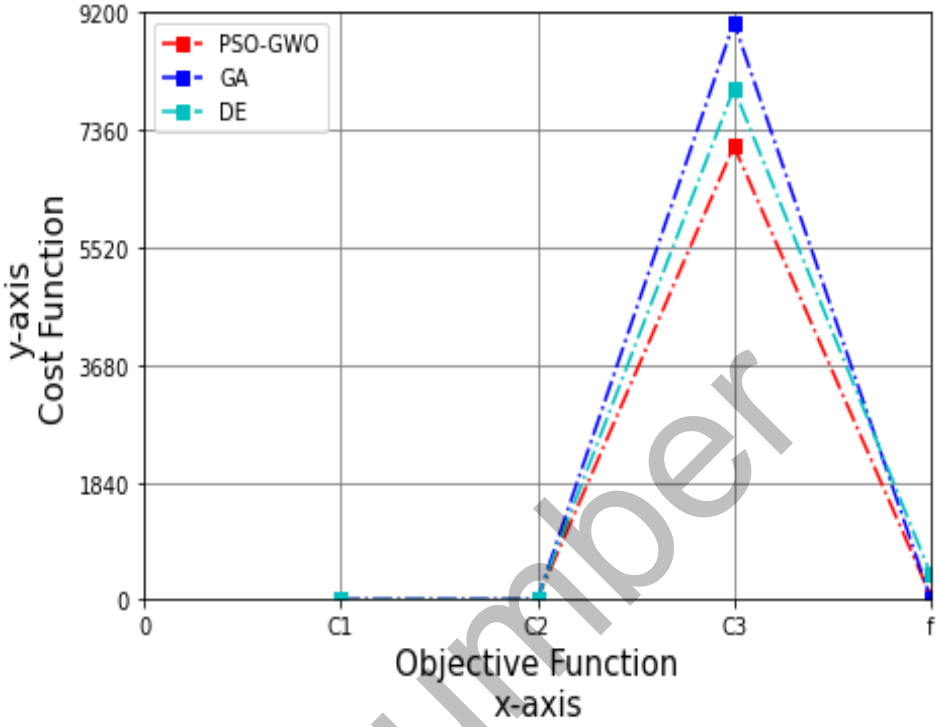



Figure 5.13 Performance analysis on cost function for autism child dataset 36 months, while $c_1 = 0.5$ and $c_2 = 0.5$.

Furthermore, the results of $C_1$, $C_2$, $C_3$ and $f$ for PSO are 0.091, 2.003, 11245.10 and 85.30, respectively; DE are 0.2, 2.009, 9289.420 and 350.609; CSA are 0.3, 2.022, 8570.42, and 10.799, and AAP-CSA are 0.4, 2.009, 9714.452, and 400.105; whereas the suggested framework achieves 0.089, 1.095, 6130.10 and 65.30, respectively, applying the 48 months autism child dataset with $c_1 = 0.5$ and $c_2 = 0.5$, which are shown in Table 5.17.

Table 5.17 Cost analysis for 48 months autism data, while $c_1 = 0.5$ and $c_2 = 0.5$

| Objective Functions | PSO | DE | CSA | AAP-CSA | PSO-GWO |
|---|---|---|---|---|---|
| $C_1$ | 0.091 | 0.2 | 0.3 | 0.4 | 0.089 |
| $C_2$ | 2.003 | 2.009 | 2.022 | 2.009 | 1.095 |
| $C_3$ | 11245.10 | 9289.420 | 8570.42 | 9714.452 | 6130.10 |
| $f$ | 85.30 | 350.609 | 10.799 | 400.105 | 65.30 |

The simulation is demonstrated in Figure 5.14, and it is revealed that this technique 2.20%, 45.33%, 45.49%, and 23.45% higher than PSO, 55.5%, 45.50%, 34%, and 81.38% greater than DE, 77.75%, 45.50%, 36.90%, and 83.68% superior to AAP-

CSA for $C_1$, $C_2$, $C_3$ and $f$, whereas 70.33%, 45.85%, and 28. 47% better than CSA for $C_1$, $C_2$, and $C_3$, respectively.

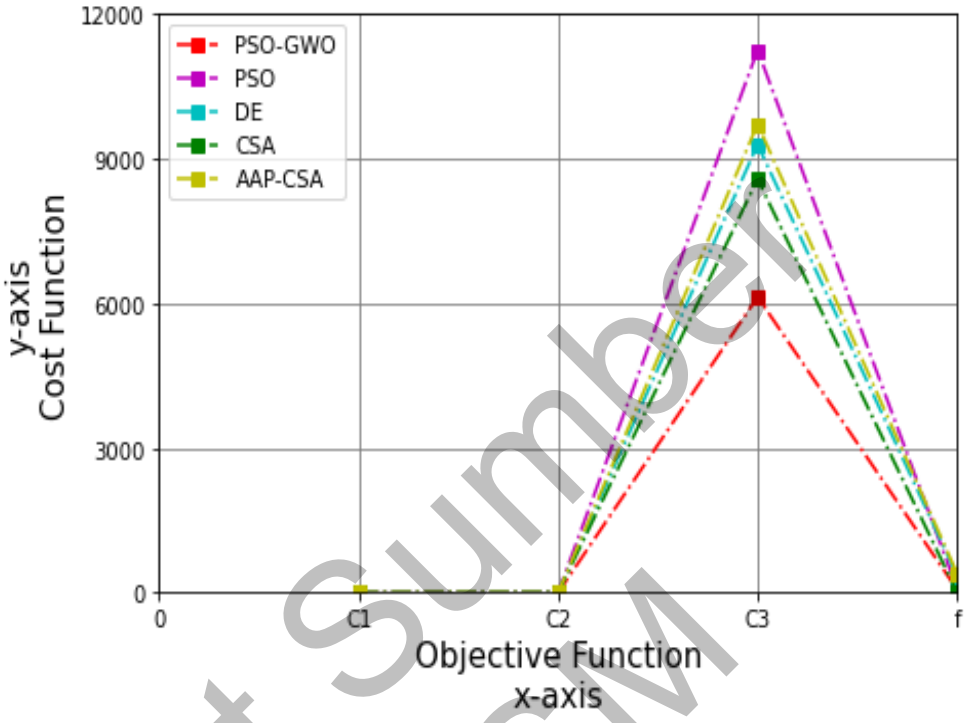



Figure 5.14 Performance analysis on cost function for autism child dataset 48 months, while $c_1 = 0.5$ and $c_2 = 0.5$.

Similarly, considering the 24 months autism child dataset to set $c_1 = 0.7$ and $c_2 = 0.7$, the outcomes of $C_1$, $C_2$, $C_3$ and $f$ for PSO are 6.731, 0.966, 430.10, and 23.10, GA are 7.333, 0.985, 2199.030 and 3.906, DE are 8.666, 0.899, 470.51, and 19.5, CSA are 2.255, 0.899, 400.002, and 0.713, AAP-CSA are 0.723, 1.002, 490.301, and 60.109, while the proposed technique achieves 0.049, 1.007, 326.25, and 24.05, respectively, that is shown in Table 5.18.

Table 5.18 Cost analysis for 24 months autism data, while $c_1 = 0.7$ and $c_2 = 0.7$

| Objective Functions | PSO | GA | DE | CSA | AAP-CSA | PSO-GWO |
|---|---|---|---|---|---|---|
| $C_1$ | 6.731 | 7.333 | 8.666 | 2.255 | 0.723 | 0.049 |
| $C_2$ | 0.966 | 0.985 | 0.899 | 0.899 | 1.002 | 1.007 |
| $C_3$ | 430.10 | 2199.030 | 470.51 | 400.002 | 490.301 | 326.25 |
| $f$ | 23.10 | 3.906 | 19.5 | 0.713 | 60.109 | 24.05 |

Figure 5.15 demonstrates that the suggested method is 99.27%, 24.15% better than PSO, 99.33%, 85.16% higher than GA, 99.43%, 30.66% superior to DE, 97.83%, 18.44% more enhanced over CSA for $C_1$, and $C_3$, whereas 93.22%, 33.46%, and 59.99% greater than AAP-CSA for $C_1$, $C_3$ and $f$, respectively.

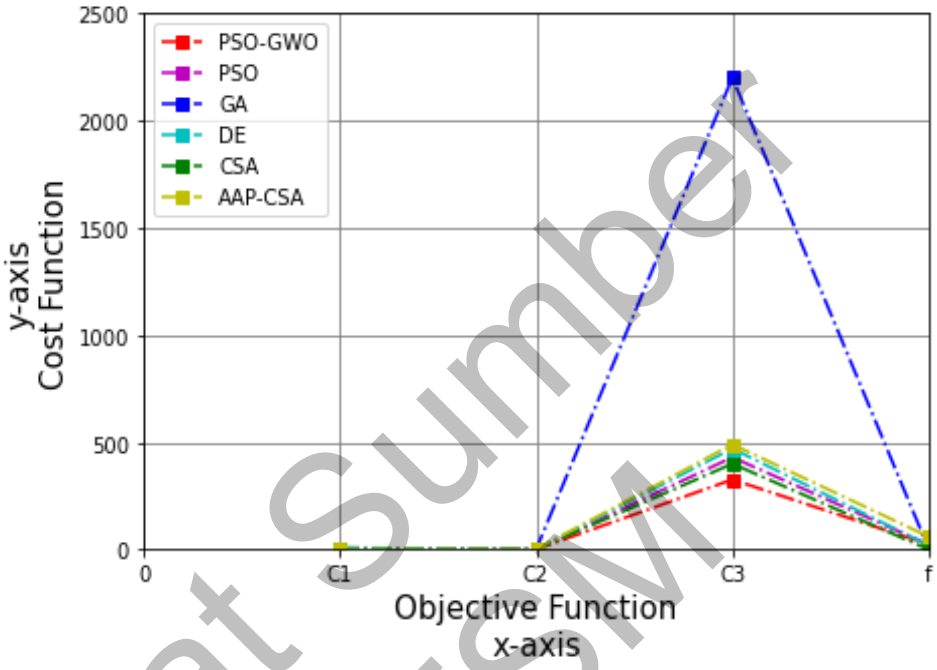



Figure 5.15 Performance analysis on cost function for autism child dataset 24 months, while $c_1 = 0.7$ and $c_2 = 0.7$.



At the values of $c_1 = 0.7$ and $c_2 = 0.7$, using 30 months autism child dataset, the results of $C_1$, $C_2$, $C_3$ and $f$ for PSO are 4.45, 0.892, 1050.35, and 20.15, GA are 2.575, 0.999, 2390.050 and 22.434, DE are 1.096, 0.898, 1310.297, and 59.501, CSA are 0.791, 2.005, 1560.998 and 4.521, AAP-CSA are 2.118, 0.899, 1055.604, and 49.608, whereas the proposed method conquers 0.005, 1.031, 950.30 and 23.15, respectively, that are shown in Table 5.19.

Table 5.19 Cost analysis for 30 months autism data, while $c_1 = 0.7$ and $c_2 = 0.7$

| Objective Functions | PSO | GA | DE | CSA | AAP-CSA | PSO-GWO |
|---|---|---|---|---|---|---|
| $C_1$ | 4.45 | 2.575 | 1.096 | 0.791 | 2.118 | 0.005 |
| $C_2$ | 0.892 | 0.999 | 0.898 | 2.005 | 0.899 | 1.031 |
| $C_3$ | 1050.35 | 2390.050 | 1310.297 | 1560.998 | 1055.604 | 950.30 |
| $f$ | 20.15 | 22.434 | 59.501 | 4.521 | 49.608 | 23.15 |

The simulation is illustrated in Figure 5.16. So, the recommended technique is 99.89%, and 9.53% greater than PSO, 99.81%, and 60.24% superior to GA for $C_1$ and $C_3$, and 99.54%, 27.47%, and 61.09% better than DE, 99.76%, 9.98%, and 53.33% more improved over AAP-CSA for $C_1$, $C_3$, and $f$, as well as 99.37%, 48.58%, and 39.12% greater than CSA for $C_1$, $C_2$, and $C_3$, respectively.

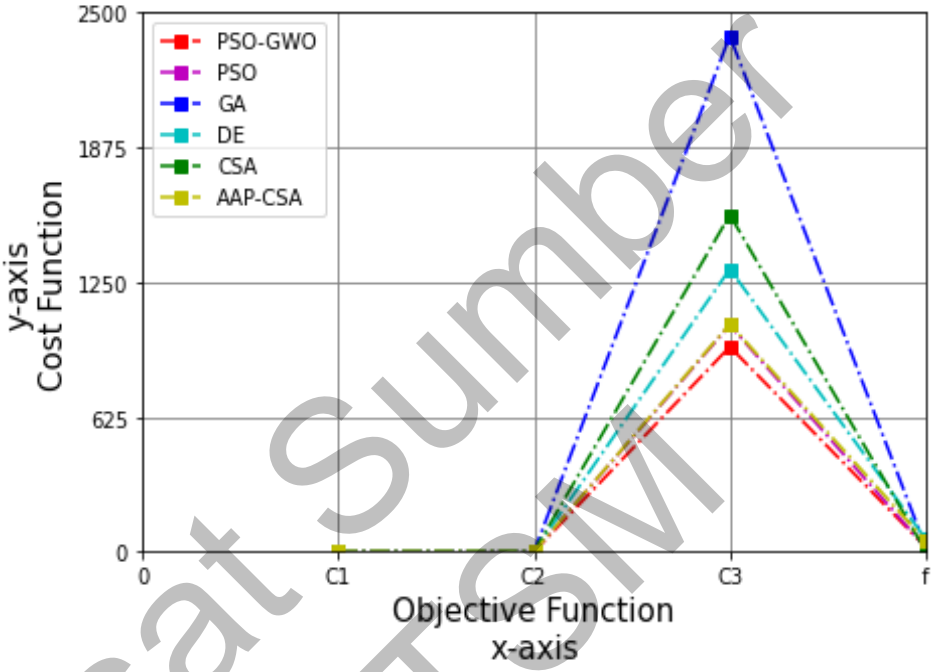



Figure 5.16 Performance analysis on cost function for autism child dataset 30 months, while $c_1 = 0.7$ and $c_2 = 0.7$.



For 36 months autism child dataset when the constant values $c_1 = 0.7$ and $c_2 = 0.7$, the results of $C_1$, $C_2$, $C_3$ and $f$ for PSO are 0.80, 2.00, 6199.15, and 130.10, GA are 0.089, 2.100, 9040.232 and 17.421, DE are 0.6, 2.112, 8001.009, and 400.706, wherein the attaining outcomes are 0.079, 1.020, 6115.15 and 45.25 correspondingly, that are presented in Table 5.20.

Table 5.20 Cost analysis for 36 months autism data, while $c_1 = 0.7$ and $c_2 = 0.7$

| Objective Functions | PSO | GA | DE | PSO-GWO |
|---|---|---|---|---|
| $C_1$ | 0.80 | 0.089 | 0.6 | 0.079 |
| $C_2$ | 2.00 | 2.100 | 2.112 | 1.020 |
| $C_3$ | 6199.15 | 9040.232 | 8001.009 | 5990.15 |
| $f$ | 130.10 | 17.421 | 400.706 | 45.25 |