An Improved Snake Optimization Algorithm for Clustering Drug Consumption

AYAD MOHAMMED JABBAR[1], HAYDER NASER KHRAIBET AL-BEHADILI[2],
KU RUHANA KU-MAHAMUD[3]
[1]Department of Translation, College of Arts,
University of Basrah,
IRAQ

[2]Department of Computer Science,
Shatt Al-Arab University,
IRAQ

[3]School of Computing,
Universiti Utara Malaysia,
MALAYSIA

Abstract: - This study used an unsupervised approach based on an optimization algorithm to cluster individuals by demographic and personality-trait similarity, and the resulting clusters were then characterized in terms of predefined drug-use categories. The phase-angle-coded snake optimization algorithm (θ-SO) has been used to solve clustering problems, but it often suffers from unsatisfactory accuracy due to sensitivity to control parameters that constrain the exploration phase. To overcome this limitation, this study proposed a snake optimization algorithm for clustering (SOC) by presenting a new equation specifically designed to improve results by directing the search to efficient regions. The optimization performance of the proposed SOC algorithm is evaluated using 19 drug consumption datasets against the most competitive algorithms reported in recent studies include genetic algorithm, grey wolf optimizer for clustering, simulated annealing and θ-SO algorithm. The experiments were conducted using four metrics: minimum clustering distance (Intra-clustering distance), purity, rand and entropy. The results were compared with mentioned algorithms indicated that the proposed SOC algorithm outperformed them by 79% based on average minimum distance, 52% based on purity, 73% based on rand and 63% based on entropy. The output from this result can be used to improve the accuracy and stability of the clustering.

Key-Words: - Unsupervised learning, Drug consumption data, Behavioral data analysis, Drug diagnosis, Metaheuristic.

Received: May 8, 2026. Revised: June 25, 2026. Accepted: July 28, 2026. Published: October 30, 2026.

1 Introduction

Contributions drug abuse in all its forms is one of the most serious threats facing humanity worldwide. Addiction is a serious and growing problem in society today. Substance abuse is a complex and multifaceted phenomenon with biological, personal, social, environmental and psychological dimensions [1]. It is not merely a behavior that threatens health, but the product of internal and external conditions and demands that play a fundamental role in regulating behavior, directing motivation and promoting social integration. This underlying meaning explains why patterns of substance abuse vary greatly amongst individuals and in different contexts. Individuals may resort to substance abuse to modify their mood or alleviate nervous tension resulting from emotional states or depression, or they may be motivated by a desire to improve their mood and engage in reckless experiences [2]. Therefore, it can be considered avoidant and a questing behavior. However, over time, this behavior can develop into a dangerous and unhealthy pattern, a growing and worsening phenomenon. Approximately 316 million people, excluding alcohol and tobacco, were addicted to drugs in 2023 [3]. Drug abuse is considered a major factor in the deterioration of drug users' health, increased crime and death rates, as well as rising unemployment and a deteriorating economic

situation [4]. The probability of an individual using drugs depends on a range of internal and external factors, defined as any characteristic, trait or event in the individual's life [5]. These factors are interconnected, such as psychological state, individual status within the family and social integration, in addition to environmental factors. However, they are primarily related to personality traits that play a key role in drug abuse.

As stated previously, the main personality traits that increase drug consumption are neuroticism, extraversion, openness to experience, agreeableness and conscientiousness. The five-factor model is the base framework used by psychologists to understand associations between traits. Studies generally indicate that neuroticism and conscientiousness are the two key elements that play a pivotal role in drug use. Although sometimes considered overlapping and sometimes contradictory, they are fundamental to drug-using behavior. Neuroticism has been associated with a predisposition to addiction, potentially linked to increased or excessive drug use, which some researchers attribute to mood instability and difficulty coping with stress [6]. Conversely, low conscientiousness characterized by reduced self-discipline and disregard for social norms has been linked to increased drug use in some studies to increased drug use [7]. Some researchers further suggest a bidirectional relationship, whereby sustained substance use may itself contribute to further declines in self-control [8], though this relationship remains debated. Moreover, some studies report an association between higher extraversion and alcohol use, potentially linked to traits such as impulsivity and increased social engagement, which may in turn correlate with a greater likelihood of experimenting with other substances such as marijuana [9]. Studies have also indicated that adaptability plays a protective role. Decreased adaptability leads to increased drug use, whereas greater adaptability reduces drug use, especially alcohol and nicotine. Given the range of factors influencing drug use, it is best to identify them using artificial intelligence (AI), due to its ability to analyze data and extract classified factors that influence other variables.

2 Related Research

AI has been employed in various industrial and healthcare fields, where it has been used to detect diseases, identify common causative factors and predict their development. It has also been applied to drug use detection by identifying disorders and behavioral patterns of users compared with normal behavior. The LASSO algorithm was used to predict disorders and behavioral patterns in users through multivariate behavioral modelling [10]. In another study, fMRI was employed using methylphenidate to determine its effects on cocaine use and to detect abnormal behavior [11]. Classification and regression algorithms have been employed to detect addictive behavior, encompassing both substance-related addictions such as tobacco, alcohol, cocaine, and other drugs and behavioral addictions, such as gambling and gaming. [12]. A support vector algorithm was also applied to detect and classify addiction based on an individual's health status using single-photon emission computed tomography [13]. Artificial neural networks (ANN) were used in another study to determine an individual's future abstinence status from drugs, such as cocaine [14]. Similar to [14], ANN was proposed to predict volatile substance abuse. The algorithm was able to determine whether a person has a desire to use drugs in the future [15]. Both related studies used ANN and deep neural networks to detect and identify online drug users and dealers over the internet and Instagram. Five thousand online conversations were analyzed to identify suspicious activity and reduce this type of crime [16], whereas 2,287 active Instagram users were checked to automatically classify those who use drugs [17]. Other related work employed a real-time substance detection instated of relying on classical self-reporting [18]. The proposed work used wearable biosensors to monitor drug use through data provided by sensors. Another study in Bangladesh on drug use amongst female students examined the health problems they face, in addition to psychological and academic factors. The study, which included 115 female students, revealed a high rate of drug use of all types, particularly amphetamines, cannabis and tranquilizers. It also showed a decline in academic performance, an increase in depression and a decrease in university attendance [19]. A study conducted in Nigeria, involving 100 football players, showed that performance pressures, the desire to recover quickly from injuries and inconsistent performance levels all contribute to problems, including psychological issues, poor concentration, anxiety and financial difficulties. These factors significantly increase the likelihood of drug use [20]. Other related work studied opioid use disorder through similarity network-based deep learning (SINDEL) to find appropriate treatment for opioid use disorder [21]. The algorithm used social media as a source for analyzing user-generated content as a rich and constantly changing dataset for studying opioid use disorder.

A similar study proposed using a convolutional neural network (CNN) to analyze speech signals and identify drug addiction [22]. This algorithm can handle speech of any length, dividing it into groups of shapes to achieve the required classification. From the studies mentioned above, we observe that personality traits and behavioral factors increase the risk of drug use of all types. The five-factor model is fundamental for psychologists, as it helps them understand the various relationships between traits and their mutual influence. Previous studies generally indicate that neuroticism and conscientiousness are the two main elements that play a pivotal role in drug use. Numerous studies have been conducted to predict when drugs will be used and their effects on different segments of society. Majority of these studies used classical supervised learning classification algorithms, with the absence of optimization algorithms and unsupervised machine learning-based clustering. This type of clustering is required due to overlapping factors amongst drug users, which makes supervised learning-based classification systems unable to handle them effectively.

Clustering techniques can be applied to organize data into different groups. Similarities and differences depend on a specific metric, such as whether the data are numerical or textual. The main drawback of the traditional approach is the difficulty in avoiding local optima. As this approach is limited in its ability to access and avoid optima, researchers are focusing on using other tools in meta-heuristic reasoning, which is inspired by natural behavior. Meta-heuristic reasoning uses a completely different clustering method and treats the search space differently, framing the clustering problem as an optimization problem. The optimization algorithm is able to find more optimal results than classical algorithms when the target is to improve several solutions simultaneously. One best behavior used in nature as inspiration for artificial intelligence optimization algorithms can be seen in snakes. Snakes are cold-blooded reptiles, characterized by their length, leglessness and exceptional speed, allowing them to move nimbly from place to place. They possess a unique ability to swallow prey larger than themselves, owing to their articulated skulls. What distinguishes snakes from other reptiles is the unique mating behavior of females, which can be seasonal or polyploid, and their reproductive patterns. They are able to control mate selection, sperm production and other aspects such as regulating body temperature and choosing preferred nesting nests. This behavior is evident in the mating process, which occurs under ideal conditions of low temperatures in late spring and early summer. However, snakes favor two main factors for mating: low temperatures and an abundance of food. These factors attract males, who then compete for the female's attention. Ultimately, the female has the choice to mate or reject the male. During mating, the female lays a clutch of eggs and waits for them to hatch, then leaves the area. This distinctive behavior in artificial intelligence algorithms can be described in two phases: exploration and exploitation. Exploration represents the absence of environmental factors such as cold and abundant food, where snakes focus exclusively on foraging in their immediate surroundings. Another factor is the presence of favorable climatic conditions, such as cold, but with scarce food which is known as exploitation. This phase involves several transitional stages to make the ecosystem more efficient. If food is plentiful but accompanied by unfavorable climatic conditions (high temperatures), snakes will focus exclusively on consuming the available food. Finally, if all conditions are satisfied, including suitable food and weather, snakes will focus on reproduction through mating. Here, two important scenarios are noted: fighting between males and mating. During fighting, males compete for the best female, and females are free to choose the best male to mate with.

The snake characteristics inspired the researchers to produce one of the best optimization algorithms that showed promising results, namely, the snake optimizer (θ-SO) algorithm for clustering [23] [24]. It has shown strong performance in solving nonlinear, high-dimensional and geometric optimization problems. However, despite these advantages, the algorithm faces certain challenges, such as the ease of falling into local optimum solutions and suboptimal convergence performance [24], Unlike the Ant Colony Optimization (ACO) algorithm, which tends to stagnate in a local optimum due to pheromone buildup on a single dominant pathway causing almost all ants to converge on the same path and requiring a restart of the search process [25]. The two main constraints on the SO for clustering in the exploration phase include: a fixed factor called C2, which requires optimization, and random numbers used in this phase. Both constraints lead to an early loss of diversity, causing convergence around the local optimum. To solve this problem, we can combine the algorithm with another algorithm or use Variable Neighborhood Search (VNS), which gives the algorithm a true balance between exploration and exploitation [26]. However, this increases the overall time complexity. The proposed study offers

two modifications. The first is online adaptive parameter control of these factors. The second reduces the number of parameters using a new formula. The latter is better than online adaptive parameter control because the latter itself is considered an optimization problem, where the main task is to find the optimal parameters. Moreover, online adaptive parameter control expands the search process, making the search space more complex to optimize. The study introduces a mathematical formulation that enables effective exploration of the search space without the need for parameter tuning, thereby improving the algorithm's optimization capability.

3 Mathematical Model and Algorithm

The algorithm includes several steps, illustrates as follows.

3.1 Initialization

Similar to other population-based metaheuristics such as Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), and Grey Wolf Optimizer (GWO), the θ-SO algorithm for clustering starts by initialising the population through generating groups of individuals. Each individual has a random uniform distribution in the search space, obtained using Equation 1, where I_i represents the position of the individual in the search space (centroids), and r indicates a random number between 0 and 1. The lower and upper bounds are represented by I_{max} and I_{min}, respectively [24].

$$I_i = I_{min} + r * (I_{max} - I_{min}) \quad (1)$$

Note that at this stage, the population is formed based on two equal groups of males and females. Let us assume the population consists of 100 individuals which comprises of 50 females and 50 males 50, calculated using Equation 2, where N is the total number of individuals and $N_{f,m}$ represents the male and female individuals.

$$N_{f,m} \quad N/2 \quad (2)$$

3.2 Group Evaluation

Each male and female group is evaluated to find the best male clustering solution $X_{best,m}$ and the best female clustering solution $X_{best,f}$ across all iterations, as shown in Table 1. Meanwhile, the best centroid positions in the search space, $p_{best,m,food}$ and $p_{best,f,food}$, are identified as shown in Table 2. Here, $X_{i,m}$ and $X_{i,f}$ represent the male and female clustering solutions at the current iteration, as illustrated in Table 1.

Table 1. Snake solution

Instances	1	2	3	4	5	6	7	8
Cluster label	2	1	3	2	2	3	2	1

Source: created by the authors

Table 2. Snake centroid matrix

Centroid No.	Attribute 1	Attribute 2	Attribute 3
Centroid 1	0.18	0.15	0.1
Centroid 2	2.1	0.31	1.75
Centroid 3	0.11	0.25	1.75

Source: created by the authors

Simultaneously, the centroid positions (food positions of male and female), $p_{i,m,food}$ and $p_{i,f,food}$, are found for each clustering solution at the current iteration, as shown in Table 2. The temperature, *Temp*, can be calculated based on the current iteration using Equation 3, where t represents the current iteration, and T is the maximum number of iterations.

$$Temp = exp\left(\frac{-t}{T}\right) \quad (3)$$

The food quantity, Q, can be calculated using Equation 4, where C_1 is a constant value set by the user to 0.5.

$$Q = C_1 * exp\left(\frac{-t-T}{T}\right) \quad (4)$$

3.3 Exploration Phase

Contrary to the original θ-SO algorithm, the exploration phase in this study removes the fixed C_2 and the random values used in the θ-SO algorithm during exploration. The exploration phase here is based on the approach that moves snake individuals with lower quality towards high-quality snakes, meaning towards the best snake position in the search space. Snake individuals with lower quality gradually improve during the intensive search process. To model the exploration phase, Equations 5 and 6 are used to guide the search process, where $p_{i,m,food}$ is the current male snake centroid position in the search space and $p_{i,f,food}$ is the current female snake centroid position at the current iteration. $p_{best,m,food}$ and $p_{best,f,food}$ represent the best male and female positions found thus far. For each search agent, the distance to the two best male and female snake positions found thus far is calculated using coefficient vectors C_1 and C_2:

$$D_{1,m} = \left|C_1 . p_{i,m,food} - p_{best,m,food}\right| \quad (5)$$

$$D_{2,f} = \left|C_2 . p_{\,i,f,food} - p_{best,f,food}\right| \quad (6)$$

The male and female snake positions are then updated by moving toward the best-known position, scaled by the coefficient vectors A_1 and A_2:

$$p_{i,m,food}^{t+1} = p_{i,m,food}^{t} - A_1 . D_{1,m} \quad (7)$$

$$p_{i,f,food}^{t+1} = p_{i,f,food}^{t} - A_2 . D_{2,f} \quad (8)$$

$$A_k = 2a.r_{1,k} - a, \quad C_k = 2.r_{2,k}, \quad k \in \{1,2\}$$
$$a = 2 - t.\frac{2}{T} \quad (9)$$

where t is the current iteration and T is the maximum number of iterations, and $r_{1,k}$, $r_{2,k}$ are independently sampled random numbers in the range [0, 1].

The coefficient controlling each update is drawn from a range that starts wide and shrinks steadily over the run, so early on it can push a candidate's position past the current best-known point rather than pulling it closer. That overshoot is what drives exploration the solution lands somewhere new instead of tightening around what's already been found. As the range shrinks, overshooting becomes less likely, and past the run's midpoint it becomes mathematically impossible, so every later update can only move a candidate closer to the best-known position. This gives a natural, built-in shift from broad exploration to fine-tuned exploitation, without needing a separate switch or manually tuned threshold. A second, independent random factor keeps scaling the pull toward the best position throughout the whole run, so even in the late exploitation phase some randomness remains, preventing the search from freezing on a local optimum.

3.4 Exploitation Phase

The exploitation phase is the stage where food exists with different qualities. Thus, the snakes move towards the best food centroid positions. The algorithm attempts to improves the clustering solutions in a gradual manner to achieve more optimal/ near-optimal results. This approach is performed in local food centroid positions where the neighbourhood of the best food positions in the search space, $p_{best,m,food}$ and $p_{best,f,food}$, are slightly modified to obtain better clustering solutions. The exploitation phase can be calculated using Equation 10, where X_{food} represents the position of $p_{best,m,food}$ or $p_{best,f,food}$, $p_{\,i,m/f,food}$ refers to the male or female centroid position, $Temp$ indicates the temperature at the current iteration, a_2 is a constant value equal to 2, and $rand$ is a random number between 0 and 1.

$$p_{i,m/f,food}^{t+1} = X_{food} \pm a_2 \times Temp \times rand \times \left(X_{food} - p_{i,m/f,food}^{t}\right) \quad (10)$$

3.5 Exploration Fight Mode and Mating Mode

The Fight mode and mating mode involve searching in depth within the solution space to discover better clustering solutions using the best-so-far results. The fight mode consists of $p_{\,i,m,food}$ and $p_{\,i,f,food}$, which are the current male and female snake centroid positions in the search space that require improvement. Here, a is a constant equal to 2, $rand$ is a random number between 0 and 1, and $p_{best,f,food}$ and $p_{best,m,food}$ are the best female and male centroid positions thus far. The fight mode can be calculated using Equations 11 and 12. Note that FM and FF represent the fighting ability of male and female snakes, respectively, and are calculated using Equations 13 and 14. $f_{best,f}$ and $f_{best,m}$ denote the best fitness values of the best female and male centroids, respectively, whereas $fi_{,m}$ and $fi_{,f}$ correspond to the fitness values of the male and female centroids at the current iteration. The overall process described by the pseudocode of SOC Algorithm 1 shown below.

$$p_{i,m,food}^{t+1} = p_{i,m,food}^{t} + a_2 \times FM \times rand \times \left(Q \times p_{best,f,food} - p_{i,m,food}^{t}\right) \quad (11)$$

$$p_{i,f,food}^{t+1} = p_{i,f,food}^{t} + a_2 \times FF \times rand \times \left(Q \times p_{best,m,food} - p_{i,f,food}^{t}\right) \quad (12)$$

$$FM = exp\left(\frac{-f_{best,f}}{fi_{,m}}\right) \quad (13)$$

$$FF = exp\left(\frac{-f_{best,m}}{fi_{,f}}\right) \quad (14)$$

The mating phase is governed by Equations 15 and 16, which update the male and female positions, respectively, relative to the food source. The mate attraction factors are then computed from these updated positions: Equation 17 calculates Mm using the fitness value obtained from Equation 15,

while Equation 18 calculates Mf using the fitness value obtained from Equation 16.

$$p_{i,m,food}^{t+1} = p_{i,m,food}^{t} + a_2 \times Mm \times rand \times \left(Q \times p_{i,f,food}^{t} - p_{i,m,food}^{t}\right) \quad (15)$$

$$p_{i,f,food}^{t+1} = p_{i,f,food}^{t} + a_2 \times Mf \times rand \times \left(Q \times p_{i,m,food}^{t} - p_{i,f,food}^{t}\right) \quad (16)$$

$$Mm = \exp\left(\frac{-f_{p\,i,(f),food}}{fi_{,m}}\right) \quad (17)$$

$$Mf = exp\left(\frac{-f_{p\,i,(m),food}}{fi_{,f}}\right) \quad (18)$$

Algorithm 1. High-level pseudo-code for the SOC Algorithm
Require: Dataset contains a set of instances
Output: A clustering solution contains instance are assignment into different clusters

```
Initialise all parameters (t, a2, k, n, C1)
Initialise the snake population
Evaluation of the Initialised snake population
while ( t < Max_iteration ) do
for 1 to Nf,m do
if Temp() > 0.7 then
        Exploration (Eq. 5, 9)
 else if Temp() < 0.7 then
if Rand() < 0.98 then
        Exploitation (Eq. 10)
else
if Rand() > 0.98 then
        Fight Mode (Eq. 11, 12)
else
Fight Mode (Eq. 15, 16)
end if
end if
end for
Evaluation of the snake population
end while
Print the best clustering solution
end algorithm
```

4 Problem Solution

Drug Consumption data were collected by Ellen Furman between March 2011 and March 2012 [27]. A total of 2,051 participants were recruited over a 12-month period. Responses from 166 participants were deemed unrealistic and excluded from the final dataset, as they were presumed to have been inattentive whilst answering the questions. During this period, some participants reported using a placebo for recreational purposes. This test was specifically designed to identify individuals who exaggerated their claims. Ultimately, the valid dataset consisted of 943 males and 942 females, resulting in a final sample size of 1,885 participants. Data were collected from participants who spoke various languages, with English being predominant (approximately 93.7%). Amongst them, 55.4% were from the United Kingdom, 29.5% from the United States, 4.6% from Canada, 2.9% from Australia, 0.3% from New Zealand and 1.1% from Ireland. The remaining participants, representing 6.3% of the total, were from other countries. Some participant data, such as country of residence, was withheld to protect privacy and minimise the risk of disclosure. Data were collected only within specific age groups, rather than recording the exact age of each participant. The samples included in this collection can be classified into three main groups. The first group is based on the age range between 18 and 55 years, as shown in Figure 1. The second group is based on samples collected to determine the effect of education on drug use in terms of increased or excessive use, as shown in Figure 2. The third group, shown in Figure 3, was determined based on ethnicity to determine its effect on drug use.

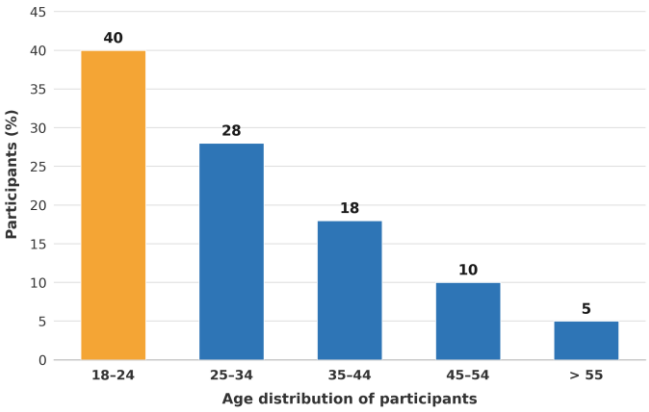


Fig. 1: Age distribution of participants (%)
Source: created by the authors

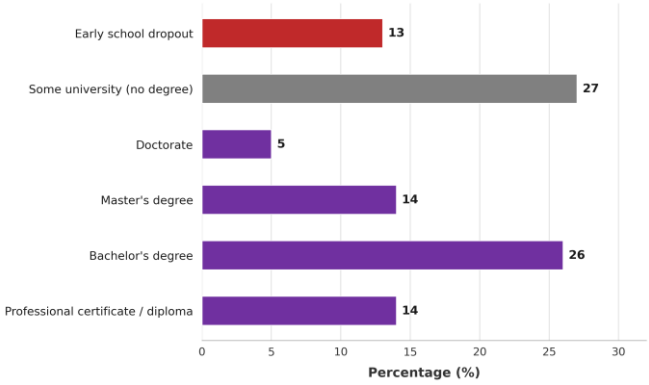


Fig. 2: Education level of drug users (%)
Source: created by the authors

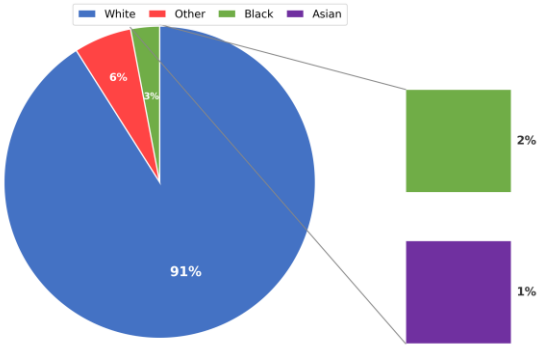


Fig. 3: Ethnic composition of participants (%)
Source: created by the authors

The collected data contains records for 1885 respondents where each record has 12 attributes used for clustering tasks, except for the ID attribute as shown in the Table 3.

Table 3. Dataset attributes

Attribute	Status	Attribute	Status
ID	Participant ID	Age	Participant Age
Gender	Participant Gender	Education	Participant Education
Country	Participant Country	Ethnicity	Participant Ethnicity
Nscore	Neuroticism Score	Escore	Extraversion Score
Oscore	Openness to Experience Score	Ascore	Agreeablene ss Score
Cscore	Conscientiou sness Score	Impulsive	Participant Impulsive
Sensation Seeing(SS)	Sensation Seeing Score		

Source: created by the authors

The dataset includes 18 legal and illegal drugs with no missing or categorical values handled includes: alcohol, amphetamine (Amphet), amyl nitrite (Amyl), benzodiazepines (Benzos), caffeine, cannabis, chocolate, cocaine, crack, ecstasy, heroin, ketamine, legal high (Legalh), lysergic acid diethylamide (LCD), methadone, magic mushrooms (Mushrooms), nicotine, volatile substance abuse (VSA) and fictitious drug (Semeron), which was introduced to identify over-claimers. The 18 drugs are shown in Appendix in Table 4, Table 5 and Table 6. For each drug, respondents must select one of the following: Never Used (CL0), Used over a Decade Ago (CL1), Used in the Last Decade (CL2), Used in the Last Year (CL3), Used in the Last Month (CL4), Used in the Last Week (CL5) and Used in the Last Day (CL6). The dataset includes multivariate characteristics and comprises 18 types of legal and illegal drugs with several destruction cases. To achieve effective clustering solutions, the dataset must first be divided into 19 independent subsets, each representing a type of drug. The number of clusters represents the dataset classes, which equals seven in each subset problem. The proposed algorithm aims to optimise each problem as an optimization problem to find the near-optimal solution. The near-optimal solution is implemented according to the objects integrated within each group, as well as by separating each group from the others.

5 Data Results and Discussion

The performance of the proposed SOC algorithm is evaluated using the 19 datasets according to the intra-clustering solution, where minimum distance guides the algorithm search. Other metrics include purity, rand and entropy. The intra-clustering solution is the total minimum error, calculated according to Equation 19, where k is the number of clusters, o_i,indicates the total number of objects within the ith cluster, and $d(x_{ij}, C_i)$ is the represents the Euclidean distance between object x_{ij} and its associated cluster centroid C_i.

$$Intra = \sum_{i=1}^{k} \sum_{j=1}^{o_i} d(x_{ij}, C_i) \qquad (19)$$

The purity metric is the number of objects accurately classified, calculated as a percentage of the total number of objects. It is included in the unit range [0...1] and can be calculated using Equation 20, as follows:

$$\text{Purity}(\Omega, \Gamma) = \frac{1}{N} \sum_{k} maxj \left| w_i \cap c_j \right| \qquad (20)$$

where $\Omega = w1, w2, \ldots, wk,$ represents the set of clusters, and $\Gamma = c1, c2, \ldots, cj,$ indicates the set of classes.

The rand metric measured using Equation 21, *RI*, which is considered the accuracy in classification, measures the percentage of decisions accurately classified. Here, TP is the true positive, FP is the false positive, FN is the false negative and TP is the true negative.

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \qquad (21)$$

The total entropy metric, *H(w)*, of all clusters can be calculated using Equation 22, which is the sum of the entropy calculated for each single

cluster. The entropy of each cluster is measured using Equation 23, as follows:

$$H(w) = -\sum_{c \in C} P(w_c) log_2 P(w_c) \quad (22)$$

where c is a classification in the set C of all classification, and $P(w_c)$ denotes the probability of a data point being classified as c in cluster w.

The total entropy of clustering is determined as follows:

$$H(\Omega) = \sum_{w \in \Omega} H(w) \frac{N_w}{N} \quad (23)$$

where $H(w)$ is a single cluster's entropy, N_w refers to the number of points in cluster w and N indicates the total number of objects.

The proposed algorithm was evaluated against the latest clustering algorithms, including the genetic algorithm (GA), the phase-angle-coded snake optimization algorithm (θ-SO), simulated annealing algorithm (SA) and the grey wolf optimization for clustering (GWOC). To ensure a fair comparison, the study followed the same set of criteria for all algorithms. In general, the number of runs is 10, the number of iterations is 1000, and the population size is 50. In the GA, the crossover rate is set to 0.8, whereas the mutation rate set to 0.001. In the SA, the probability threshold is set to 0.98, the initial temperature to 5, the temperature multiplier to 0.98 and the final temperature to 0.001. In the GWOC, the coefficient value is set to 2/NA, whereas θ-SO and SOC have C_1 set to 0.5 and a set to 2. Table 7, Table 8, Table 9 and Table 10 in Appendix present the comparative results obtained to evaluate the performance of the SOC algorithm. Table 7 (Appendix) illustrates the minimum distance clustering solution produced by all algorithms, whereas Table 8, Table 9 and Table 10 in Appendix present the purity, rand and entropy metrics obtained from the comparisons.

A comparison of the optimization algorithms shown in Table 7 (Appendix) revealed that the SOC algorithm outperformed all others. The algorithms' performance was evaluated using the average minimum distance clustering solution, which indicates how closely the elements within each group are clustered and the effective separation of the resulting clusters. The SOC algorithm produced the best result in 15 of the 19 datasets (78.9%), outperforming all other algorithms on those datasets, and outperformed the θ-SO algorithm specifically in 19 of 19 datasets (100% win rate). The GWOC algorithm achieved the best results only in two datasets, which equals 10% in this ranking.

The second comparison was conducted using the same optimization algorithms based on the purity metric results. The purity metric measures the purity of groups, indicating whether a group contains elements of the same class or type, as shown in Table 8 (Appendix). The SOC algorithm produced the best results in 10 datasets, surpassing the other algorithms by 52%, whereas the θ-SO algorithm achieved the best results in three datasets (15%) compared with the proposed algorithm. The next algorithms in this ranking were GA and GWOC, which achieved similar results in only one dataset (5%).

The third comparison, based on the rand metric, represents the accuracy of whether an element belongs to its group or not. Therefore, the metric indicates how well the set is constructed. Simply put, the index compares how well pairs of data points are grouped together in the predicted set versus the actual set. The rand metric provides a single score indicating the degree of agreement between the two sets, as shown in Table 9 (Appendix). The SOC algorithm outperformed the other algorithms in 14 datasets, which equals 73%, whereas it surpassed the θ-SO algorithm in all datasets, which equals 100%.

The final comparison, based on the entropy metric, represents the accuracy of decisions performed by the algorithm by measuring the randomness of each cluster, as shown in Table 10 (Appendix). The SOC algorithm outperformed the other algorithms in 12 datasets (63%) and surpassed the θ-SO algorithm in all datasets (100%).

Results from Table 7, Table 8, Table 9 and Table 10 in Appendix are summarised in Figure 4. This provides an overview of the algorithm's performance based on the metrics used to determine the quality of the results. The SOC algorithm has outperformed the GA algorithm by 100% in average minimum distance, 68% in purity, 94% in rand and 73% in entropy. The SOC algorithm also exceeded the SA algorithm by 100% in average minimum distance, 78% in purity and rand and 84% in entropy. A comparison between the SOC and GWOC algorithms shows that the SOC algorithm surpassed the GWOC algorithm by 100% in average minimum distance, 78% in purity and 100% in rand and entropy.

Figure 5 compares the convergence behavior of the proposed SOC mechanism against the original θ-SO mechanism, averaged over 10 independent runs, with shaded regions denoting ±1 standard deviation. The original θ-SO mechanism converges slowly and stabilizes at a mean fitness value of approximately 8.0, with a wide standard deviation band indicating

inconsistent performance across runs due to its reliance on randomly selected peer positions. In contrast, the proposed SOC mechanism converges rapidly toward the global optimum within the first 20–25 iterations, with a narrowing standard deviation band reflecting high consistency across runs. This improvement is attributed to the coefficient-based attraction mechanism, governed by the linearly decreasing parameter a, which enables a smooth transition from exploration to exploitation. These results provide empirical evidence that SOC effectively mitigates the premature convergence observed in the original θ-SO formulation.

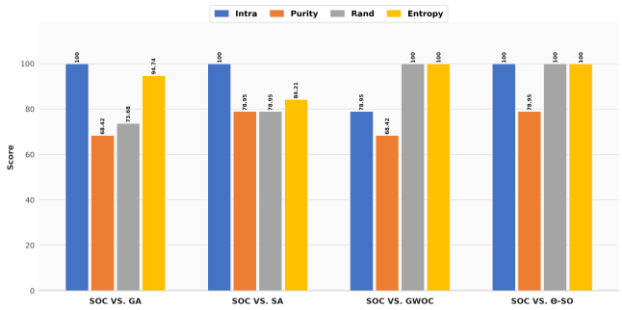


Fig. 4: Performance metrics of SOC vs. best known algorithms
Source: created by the authors

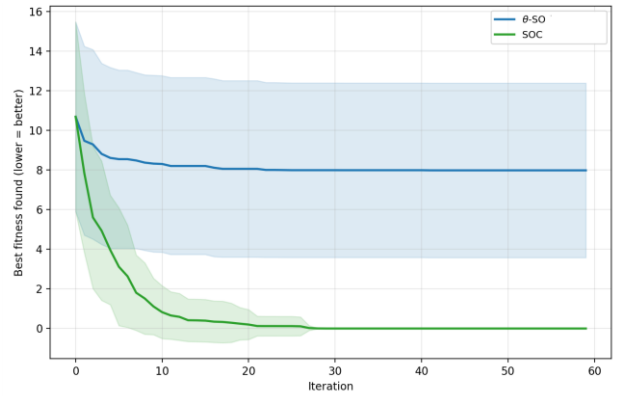


Fig. 5: The convergence behavior SOC vs. θ-SO
Source: created by the authors

6 Conclusion

This study modifies the exploration phase of SOC, where information exchange amongst the search history is limited. A new exploration formula has been proposed to direct the search process towards areas with high-quality solutions. Information exchange between high-quality and low-quality solutions allows the utilisation of search history during algorithm execution. The scientific contribution of this study lies in removing fixed and random parameters from the exploration phase of the SOC algorithm by using a stable exploration formula. The benefit is an increase in the proportion of explorations carried over to subsequent search phases. Another contribution is the ability to explore the surrounding area whilst converging towards the near-optimal solution and thereby identifying more near-optimal solution during the search process.

The proposed SOC algorithm has shown to produce better clustering results than GA, SA and GWOC algorithms in this research. The evaluation showed that the proposed algorithm achieved superior results to the related θ-SO clustering algorithm, obtaining 100% internal clustering distance, 85% in purity, 100% in rand criteria and 100% in entropy. The results demonstrate that the proposed SOC algorithm outperformed all the above-mentioned algorithms by 52% based on purity criteria, 73% based on rand criteria and 63% based on entropy criteria.

Further work can focus on the exploitation phase which requires modification due to its heavy reliance on random disturbances and the absence of strong local optimization mechanisms based on reinforcement information. This issue can be solved using local research methods such as hybridisation with hill climbing and the simple Nelder–Mead simplex search. Another area that can be researched is the application of the algorithm directly to similar problems, such as the vehicle routing problem and the traveling salesman problem, by integrating in-depth search strategies with adaptive parameters to reduce sensitivity to random parameters

Acknowledgments:
The authors would like to express their sincere gratitude to University of Basrah and Shatt Al-Arab University for supporting this research and providing research facilities and laboratory resources.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work, the author(s) utilized artificial intelligence (AI) technologies solely for the purpose of linguistic refinement and grammatical correction. Following the application of these tools, the author(s) comprehensively reviewed, verified, and edited the resulting text. The author(s) maintain full accountability and responsibility for the accuracy, integrity, and final content of the publication.

References:

[1] B. Ahmed, F. Navid Yousaf, M. Saud, and A. Ahmad, “Youth at risk: The alarming

issue of drug addiction in academic institutions in Pakistan," *Child. Youth Serv. Rev.*, vol. 118, Nov. 2020, doi: 10.1016/j.childyouth.2020.105385.

[2] L. M. Matei-Mitacu, T. D. Huțul, A. Karner-Huțuleac, A. Huțul, and C. A. Dobria, "The role of alcohol consumption motives in the relationships between psychological distress, emotional dysregulation, and problematic alcohol consumption. A mediation model," *Current Psychology*, vol. 43, no. 48, pp. 36831–36845, Dec. 2024, doi: 10.1007/s12144-024-07111-0.

[3] U. N. publication UNODC, "WORLD DRUG REPORT 2025," United Nations Office on Drugs and Crime's post, Vienna, 2025. Accessed: Feb. 06, 2026. [Online]. Available: https://www.unodc.org/documents/data-and-analysis/WDR_2025/WDR25_B1_Key_findings.pdf

[4] D. A. Glei and M. Weinstein, "Drug and alcohol abuse: The role of economic insecurity," *Am. J. Health Behav.*, vol. 43, no. 4, pp. 838–857, 2019, doi: 10.5993/AJHB.43.4.16.

[5] E. Martínez-Gabaldón and J. E. Martínez-Pérez, "Personality Traits and Drug Use: A Longitudinal Study Using Data from the British Cohort Study," *Eur. Addict. Res.*, vol. 30, no. 6, pp. 357–365, Mar. 2025, doi: 10.1159/000541265.

[6] S. Gong, J. Dong, L. Wang, X. Zhao, L. Zheng, and X. Xi, "Significance of constructing quality of care sensitive indicators based on Donabedian structure-process-outcome three-dimensional quality model in quality control of pediatric endoscopy," *Rev. Psiquiatr. Clín.*, vol. 49, no. 6, pp. 55–60, 2022, doi: 10.15761/0101-60830000000498.

[7] J. C. Thennakoon, D. D. Adithiya, and A. R. Jayamaha, "The Association Between Personality Traits and Substance Use Among Advanced Level Students in Western Province, Sri Lanka: A Cross-sectional Study," *Addiction and Health*, vol. 16, no. 2, pp. 100–106, May 2024, doi: 10.34172/ahj.2024.1520.

[8] C. M. Juchem, A. Bendau, L. C. Bandurski, N. J. Reich, S. Baumgardt, and E. Asselmann, "Personality changes related to presence and treatment of substance use (disorders): a systematic review," Jul. 01, 2024, *Cambridge University Press*. doi: 10.1017/S003329172400093X.

[9] I. H. A. Franken and P. Prinzie, "Adolescent extraversion and agreeableness predict adult alcohol use: A 22-Year prospective study," *Addictive Behaviors*, vol. 165, Jun. 2025, doi: 10.1016/j.addbeh.2025.108303.

[10] J. J. Prisciandaro, A. L. McRae-Clark, H. Myrick, S. Henderson, and K. T. Brady, "Brain activation to cocaine cues and motivation/treatment status," *Addiction Biology*, vol. 19, no. 2, pp. 240–249, Mar. 2014, doi: https://doi.org/10.1111/j.1369-1600.2012.00446.x.

[11] M. J. McHugh, H. Gu, Y. Yang, B. Adinoff, and E. A. Stein, "Executive control network connectivity strength protects against relapse to cocaine use," *Addiction Biology*, vol. 22, no. 6, pp. 1790–1801, Nov. 2017, doi: https://doi.org/10.1111/adb.12448.

[12] X. Liang, Y. He, B. J. Salmeron, H. Gu, E. A. Stein, and Y. Yang, "Interactions between the salience and default-mode networks are disrupted in cocaine addiction," *Journal of Neuroscience*, vol. 35, no. 21, pp. 8081–8090, May 2015, doi: 10.1523/JNEUROSCI.3188-14.2015.

[13] W. Y. Ahn, D. Ramesh, F. G. Moeller, and J. Vassileva, "Utility of machine-learning approaches to identify behavioral markers for substance use disorders: Impulsivity dimensions as predictors of current cocaine dependence," *Front. Psychiatry*, vol. 7, no. MAR, Mar. 2016, doi: 10.3389/fpsyt.2016.00034.

[14] M. Mete, U. Sakoglu, J. S. Spence, M. D. Devous, T. S. Harris, and B. Adinoff, "Successful classification of cocaine dependence using brain imaging: A generalizable machine learning approach," *BMC Bioinformatics*, vol. 17, Oct. 2016, doi: 10.1186/s12859-016-1218-z.

[15] Priyanka Nath, Sumran Kilam, and Aleena Swetapadma, "A Machine Learning Approach to Predict Volatile Substance Abuse for Drug Risk Analysis," in *2017 Third International Conference on Research in Computational Intelligence and Communication Networks*, IEEE, 2017.

[16] X. Yang and M. Ding, "Identification of drug-related chat content based on artificial neural network," in *Proceedings - 2020 12th International Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2020*, Institute of Electrical and

Electronics Engineers Inc., Feb. 2020, pp. 526–530. doi: 10.1109/ICMTMA50254.2020.00119.

[17] S. Hassanpour, N. Tomita, T. DeLise, B. Crosier, and L. A. Marsch, "Identifying substance use risk based on deep neural networks and Instagram social media data," *Neuropsychopharmacology*, vol. 44, no. 3, pp. 487–494, Feb. 2019, doi: 10.1038/s41386-018-0247-x.

[18] Jin Wang, Hua Fang, Stephanie Carreiro, Honggang Wang, and Edward Boyer, "A New Mining Method to Detect Real Time Substance Use Events from Wearable Biosensor Data Stream," in *2017 International Conference on Computing, Networking and Communications (ICNC)*, USA: IEEE, Jan. 2017, pp. 465–470.

[19] F. Akter *et al.*, "Evaluation of Patterns and Risk Factors of Substance Abuse among Female Students in Dhaka City, Bangladesh," *Community Based Medical Journal*, vol. 15, no. 1, pp. 216–223, Feb. 2026, doi: 10.3329/cbmj.v15i1.87641.

[20] U.-A. Paul Aniekanobong, O. Adebayo Ganiyu, A. Joshua Anojie, S. Olabode Idris, and K. Sodeeq Adekunle, "Effects of substance abuse on the mental health of football players in Lagos State, Nigeria," *Akasa Multidisciplinary Journal*, pp. 332–348, 2026, [Online]. Available: https://akasa.com.ng

[21] J. Xie, Z. Zhang, X. Liu, and D. Zeng, "Unveiling the Hidden Truth of Drug Addiction: A Social Media Approach Using Similarity Network-Based Deep Learning," *Journal of Management Information Systems*, vol. 38, no. 1, pp. 166–195, 2021, doi: 10.1080/07421222.2021.1870388.

[22] A. H. Hadjahmadi, M. H. Ziyaaddini, and M. M. Omid, "Using End-to-End Deep Neural Networks for Diagnosing Drug Addiction from Speech Signals," in *6th Iranian Conference on Signal Processing and Intelligent Systems, ICSPIS 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020. doi: 10.1109/ICSPIS51611.2020.9349595.

[23] F. A. Hashim and A. G. Hussien, "Snake Optimizer: A novel meta-heuristic optimization algorithm," *Knowl. Based. Syst.*, vol. 242, Apr. 2022, doi: 10.1016/j.knosys.2022.108320.

[24] D. Xue, S. Y. Pang, N. Liu, S. K. Liu, and W. M. Zheng, "Phase-Angle-Encoded Snake Optimization Algorithm for K-Means Clustering," *Electronics (Switzerland)*, vol. 13, no. 21, Nov. 2024, doi: 10.3390/electronics13214215.

[25] A. A. -, K. R. K. -, and N. Md. N. -, "Interacted Multiple Ant Colonies Optimization Framework: an Experimental Study of the Evaluation and the Exploration Techniques to Control the Search Stagnation," *Int. J. Adv. Comput. Technol.*, vol. 2, no. 1, pp. 78–85, Mar. 2010, doi: 10.4156/ijact.vol2.issue1.7.

[26] H. N. K. Al-Behadili, "Improved firefly algorithm with variable neighborhood search for data clustering," *Baghdad Science Journal*, vol. 19, no. 2, pp. 409–421, 2022, doi: 10.21123/BSJ.2022.19.2.0409.

[27] E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan, and A. N. Gorban, "The five factor model of personality and evaluation of drug consumption risk," in *Studies in Classification, Data Analysis, and Knowledge Organization*, Springer Science and Business Media Deutschland GmbH, 2017, pp. 231–242. doi: 10.1007/978-3-319-55723-6_18.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0
https://creativecommons.org/licenses/by/4.0/deed.en_US

APPENDIX

Table 4. Drug consumption (Part One)

Class	Alcohol	Amphet	Amyl	Benzos	Caffeine	Cannabis
CL0	34	976	1305	1000	27	413
CL1	34	230	210	116	10	207
CL2	68	243	237	234	24	266
CL3	198	198	92	236	60	211
CL4	287	75	24	120	106	140
CL5	759	61	14	84	273	185
CL6	505	102	3	95	1385	463

Source: created by the authors

Table 5. Drug consumption (Part two)

Class	Chocolate	Cocaine	Crack	Ecstasy	Heroin	Ketamine	Legal high
CL0	32	1038	1627	1021	1605	1490	1094
CL1	3	160	67	113	68	45	29
CL2	10	270	112	234	94	142	198
CL3	54	258	59	277	65	129	323
CL4	296	99	9	156	24	42	110
CL5	683	41	9	63	16	33	64
CL6	807	19	2	21	13	4	67

Source: created by the authors

Table 6. Drug consumption (Part three)

Class	LSD	Methadone	Mushrooms	Nicotine	Semer	VSA
CL0	1069	1429	982	428	1877	1455
CL1	259	39	209	193	2	200
CL2	177	97	260	204	3	135
CL3	214	149	275	185	2	61
CL4	97	50	115	108	1	13
CL5	56	48	40	157	0	14
CL6	13	73	4	610	0	7

Source: created by the authors

Table 7. Average minimum distance clustering solution: GA Vs. SA Vs. GWOC Vs. Θ-SO Vs. SOC

Dataset	GA	SA	GWOC	θ-SO	SOC
Alcohol	5523.4780	5857.8040	5045.1441	5116.3015	**4963.5986**
Amphet	5370.9834	5683.4990	**4814.1658**	5032.10442	4894.1508
Amyl	5227.7514	5518.0536	**4690.1964**	4958.0387	4772.6041
Benzos	5492.0726	5721.3714	5176.8380	5027.325	**4907.1997**
Caffeine	5232.4494	5506.6168	4990.1358	4955.7924	**4777.3306**
Cannabis	5659.4691	6095.6587	5136.3577	5195.4798	**4990.1075**
Chocolate	5220.1947	5477.3172	4989.1320	4947.7893	**4816.2218**
Cocaine	5475.0750	5805.1596	5185.6868	5084.6333	**4928.8613**
Crack	5103.5825	5376.6046	4783.0476	4824.4639	**4705.3923**
Ecstasy	5478.7209	5854.8949	5173.3307	5057.8073	**4954.5086**
Heroin	5167.0325	5469.2377	4878.9907	4866.2215	**4765.3811**
Ketamine	5210.9725	5513.5310	4939.4073	4898.8966	**4781.2418**
Legal high	5445.2455	5782.1708	5089.7953	5042.5618	**4916.2403**
LSD	5363.3655	5659.4785	5037.2035	5017.2201	**4860.2448**
Methadone	5335.1939	5643.9051	5057.1339	4923.5507	**4808.6593**
Mushrooms	5338.0817	5666.5822	5033.9574	4972.3908	**4874.2737**
Nicotine	5577.5371	5911.5501	5300.2488	5216.6568	**5050.6423**
Semer	4909.7486	5175.6620	4540.3722	4618.8437	4540.3722
VSA	5311.3604	5511.2887	4756.5787	4878.3597	4756.5787

Source: created by the authors

Table 8. Average purity-clustering solution: GA Vs. SA Vs. GWOC Vs. Θ-SO Vs. SOC

Dataset	**GA**	**SA**	**GWOC**	**θ-SO**	**SOC**
Alcohol	0.5655	0.7163	0.6786	0.6786	**0.7917**
Amphet	0.7917	0.5655	0.5655	0.5655	**0.8294**
Amyl	0.5655	0.6032	0.5655	0.6032	**0.9802**
Benzos	**0.7917**	0.5655	0.1885	0.5278	0.7163
Caffeine	0.3770	0.5655	**0.7917**	0.5278	0.6786
Cannabis	0.5655	0.5655	0.5655	**0.6032**	0.5655
Chocolate	0.7917	0.9048	0.5655	**0.9802**	0.7917
Cocaine	0.7163	0.5655	0.3770	0.5655	**0.7917**
Crack	0.7917	0.5655	0.7917	0.5655	0.7917
Ecstasy	0.5655	0.5655	0.5655	**0.7163**	0.6409
Heroin	0.5655	0.5655	0.7917	0.6786	0.7917
Ketamine	0.7917	0.7917	0.7917	0.7917	0.7917
Legal high	0.7917	0.7917	0.7917	0.5655	0.7917
LSD	0.5655	0.5655	0.5655	0.5278	**0.6409**
Methadone	0.7917	0.7917	0.5655	0.5655	**0.8671**
Mushrooms	0.5655	0.7917	0.5655	0.5655	**0.8294**
Nicotine	0.7917	0.6786	0.3770	0.6409	**0.9802**
Semer	0.5655	0.6786	0.5655	0.6409	**0.7917**
VSA	0.5655	0.7540	0.5655	0.5655	**0.8671**

Source: created by the authors

Table 9. Average rand-clustering solution: GA Vs. SA Vs. GWOC Vs. Θ-SO Vs. SOC

Dataset	**GA**	**SA**	**GWOC**	**θ-SO**	**SOC**
Alcohol	0.8560	0.8561	0.8439	0.8201	0**.8568**
Amphet	0.8568	0.8443	0.8505	0.8154	**0.8573**
Amyl	0.8567	0.8434	0.8521	0.8172	**0.8573**
Benzos	0.8377	0.8563	0.8232	0.8192	**0.8570**
Caffeine	0.8569	0.8309	0.8387	0.7872	**0.8572**
Cannabis	0.8559	0.8569	0.8250	0.8312	0.8569
Chocolate	0.8559	0.8567	0.8422	0.8145	0.8567
Cocaine	0.8569	0.8547	0.8159	0.8239	0.8569
Crack	0.8566	0.8371	0.8385	0.7931	**0.8571**
Ecstasy	0.8558	0.8572	0.7968	0.8288	0.8572
Heroin	0.8567	0.8572	0.8486	0.8114	0.8572
Ketamine	0.8559	0.8370	0.8489	0.7993	**0.8570**
Legal high	0.8560	0.8290	0.8301	0.8142	**0.8571**
LSD	0.8550	0.8460	0.8385	0.8120	**0.8573**
Methadone	0.8557	0.8233	0.8171	0.8025	**0.8564**
Mushrooms	0.8544	0.8447	0.8178	0.8106	**0.8573**
Nicotine	0.8567	0.8525	0.8460	0.8418	**0.8572**
Semer	0.8567	0.8550	0.8424	0.8230	**0.8573**
VSA	0.8543	0.8547	0.7799	0.8065	**0.8564**

Source: created by the authors

Table 10. Average entropy-clustering solution: GA Vs. SA Vs. GWOC Vs. Θ-SO Vs. SOC

Dataset	**GA**	**SA**	**GWOC**	**θ-SO**	**SOC**
Alcohol	0.0255	0.0257	0.0257	0.0269	**0.0254**
Amphet	0.0259	0.0268	0.0264	0.0275	**0.0256**
Amyl	0.0258	0.0266	0.0264	0.0270	**0.0256**
Benzos	0.0259	0.0273	0.0284	0.0278	**0.0255**
Caffeine	0.0256	0.0256	0.0273	0.0269	0.0253
Cannabis	0.0254	0.0256	0.0267	0.0267	0.0254
Chocolate	0.0255	0.0258	0.0265	0.0265	0.0255
Cocaine	0.0277	0.0257	0.0283	0.0270	**0.0256**
Crack	0.0271	0.0259	0.0261	0.0277	**0.0257**
Ecstasy	0.0255	0.0256	0.0278	0.0270	0.0255
Heroin	0.0259	0.0255	0.0261	0.0263	**0.0251**
Ketamine	0.0257	0.0255	0.0271	0.0273	**0.0254**
Legal high	0.0259	0.0256	0.0244	0.0271	**0.0241**
LSD	0.0255	0.0255	0.0260	0.0275	0.0255
Methadone	0.0258	0.0258	0.0278	0.0276	0.0258
Mushrooms	0.0263	0.0257	0.0262	0.0274	0.0257
Nicotine	0.0259	0.0255	0.0261	0.0257	**0.0254**
Semer	0.0256	0.0263	0.0264	0.0268	**0.0255**
VSA	0.0271	0.0260	0.0284	0.0272	**0.0259**

Source: created by the authors