Reptile Search Algorithm Based Feature Selection Approach for Intrusion Detection

Maher O Al-Khateeb[1] and Ali Douik[2]
[1]*ISITCom, NOCCS-ENISO Lab, University of Sousse, Sousse, Tunisia*
[2]*National Engineering School of Sousse, NOCCS-ENISO Lab, University of Sousse, Sousse, Tunisia*

Keywords: Reptile Search Algorithm, Feature Selection, Intrusion Detection, Cybersecurity, Metaheuristic.

Abstract: In Cybersecurity, the Rise of Machine Learning (ML) Based Security Solutions Has Led to a New Era of Defense Against Evolving Threats, with Intrusion Detection (ID) Systems at the Forefront. However, the Effectiveness of These Systems Is Profoundly Influenced by the Quality and Relevance of the Input Features. the Presence of Redundant Features Can Compromise Their Performance, Making Feature Selection (FS) a Crucial Step in Optimizing ID Solutions. This Paper Uses the Reptile Search Algorithm (RSA) as a Powerful FS Method. It Offers a Gradient-Free Approach, Avoiding Local Optima and Enabling Global Optimization. Comparative Analysis Using Five Freely Available ID Datasets and Benchmarked Against Several Methods Validated Superior Performance of the RSA for ID.

1 INTRODUCTION

The deployment of Internet of Things (IoT), information technology and operational environments have given rise to new cybersecurity risks. These risks threaten the security of operational ecosystems, safety, and efficiency, posing a danger to physical and financial wellbeing (Yadav,2023). The growth of cyber-attacks threat affects businesses, social networks, digital privacy, and precarious infrastructure. ID systems play a crucial role in enhancing the security of IT infrastructures. They are effective in detecting and countering attacks, providing protection against intrusive hackers (Khan,2020).

An intrusion is characterized as unexpected activities that can harm the confidentiality, integrity, and availability of the network. To detect anomalies, IDs analyze network traffic and packet header fields to identify unusual patterns, thereby preventing or minimizing damage to the network or system (Alsoufi,2021). The primary goal of an IDs is to identify and avert unauthorized use and both any kind of network intrusions, hence boosting the overall security of the network.

IDs are typically deployed on network nodes or hosts and use a combination of signature-based and anomaly-based detection techniques. Signature-based detection involves comparing network traffic or system activity against a database of known attack patterns or signatures (Khraisat,2019). Anomaly-based detection involves analysing network traffic or system activity to identify behaviours that deviate from normal patterns. IDs can be classified into two types: Network-based Intrusion Detection Systems (NIDS) and Host-based Intrusion Detection Systems (HIDS). NIDS analyse network traffic and look for patterns that indicate an intrusion attempt, while HIDS analyse activity on individual hosts, such as system calls and file access patterns.

ML is a subset of artificial intelligence that involves training algorithms to analyse and learn from data. ML algorithms can be used to classify data, make predictions, and identify patterns in large datasets [(Liu,2019), (Al-Khateeb,2021)]. These techniques are particularly useful for solving problems, where the solution is not well-defined or where there may be a large number of variables to consider. FS methods aim to exclude features that are unrelated and redundant while retaining the salient ones. This process not only enhances overall performance but also reduces data dimensionality, resulting in a lower cost of classification by decreasing the training time required to build less complex ML models (Al-Shourbaji,2023). On the other hand, using all features in the model increases computational overhead, training and testing times, storage requirements, and

Al-Khateeb, M. O. and Douik, A.
Reptile Search Algorithm Based Feature Selection Approach for Intrusion Detection.
DOI: 10.5220/0013166800003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 639-645
ISBN: 978-989-758-737-5; ISSN: 2184-433X

error rate of ML model due to irrelevant features confusing with the relevant ones.

Metaheuristic (MH) algorithm, also known as a MH optimization algorithm, is a general FS algorithmic framework that can be used to find optimal solutions in a wide range of problem domains (Fong,2016). These algorithms are designed to solve complex optimization problems, where traditional approaches may be insufficient. They are typically inspired by natural processes like evolution, swarm behaviour, and other complex systems (Xu,2014). They use these models to develop search strategies that can efficiently navigate complex search spaces, avoiding local optima and finding globally optimal solutions. One of the key advantages of MH algorithms is that they are very flexible and can be adapted to solve a wide range of problems. They are also often faster and more efficient than traditional optimization techniques, making them an attractive option for large-scale optimization problems.

Some popular examples of MH algorithms include, genetic algorithms, Particle Swarm Optimization (PSO) (Kennedy,1995), Grey Wolf Optimization (GWO) (Mirjalili,2014), Multi-Verse Optimizer (MVO) (Mirjalili,2016), Remora Optimization Algorithm (ROA) (Jia,2021), genetic algorithm (Holland,1992) and many others. These algorithms have been successfully applied to a diverse range of fields, including engineering, finance, operations research, and many others. Recently, Reptile Search Algorithm (RSA) (Abualigah,2022), shows a great potential as a FS method and it can pick Optimal Feature Subset (OFS) effectively. This paper aims to investigate FS method using RSA for ID. To assess RSA's capabilities in determining OFS, five publicly available ID datasets and various quantitative evaluation measures are used. Four FS methods, PSO, GWO, MVO, and ROA, are implemented to compare RSA's efficiency in ID system.

The organization of remaining paper is as follows: Section 2, briefly describes the RSA and datasets used. Section 3, describes evaluations measurements to evaluate RSA method. Section 4 discusses the experimental analysis and results. Section 5 concludes the paper.

2 METHODS AND MATERIALS

2.1 Reptile Search Algorithm (RSA)

RSA is a new method inspired by the hunting behaviour of Crocodiles proposed by (Abualigah,2022) in 2022. A set of candidate N crocodiles $x_{i,j}$ each having random position in the search space are initialized as follows:

$$x_{i,j} = rand_{\in U(0,1)} * \left(UB_j - LB_j\right) + LB_j \quad i \in \{1, \dots, N\} \; and \; j \in \{1, \dots, M\} \tag{1}$$

where LB_j and UB_j are the lowest and highest values of the jth feature, $rand_{\in U(0,1)}$ generates a number randomly in the range [0, 1] following a uniform distribution, and M is feature dimensionality in the dataset.

For crocodile food search, two distinct strategies, exploration and exploitation, are employed. These strategies are sequentially implemented over four stages within the maximum iteration limit. In the initial half of these stages, the algorithm leverages the crocodile's encircling behaviour, incorporating both high and belly walking movements, to facilitate search space exploration. It can be formulated as:

$$x_{i,j}(g+1) = \begin{cases} \left[-n_{i,j}(g) \,.\, \gamma \,.\, Best_j(g)\right] - \left[rand_{\in[1,N]} \,.\, R_{i,j}(g)\right], & g \leq \frac{T}{4} \\ ES(g) \,.\, Best_j(g) \,.\, x_{(rand_{\in[1,N]},j)}, & g \leq \frac{2T}{4} \; and \; g > \frac{T}{4} \end{cases} \tag{2}$$

where, for gth iteration, the best position for gth feature is $Best_j(g)$, the hunting operator $n_{i,j}$ is calculated as in Eq. (3), and the parameter for Evolutionary Sense $ES(g)$ is calculated as in Eq. (7). ES parameter decreases over the iterations between 2 to −2. Finally, the exploration accuracy is controlled by setting parameter γ as 0.1. The search region is continuously decreased by parameter $R_{i,j}$, calculated as in Eq. (6). A crocodile is randomly selected by $rand_{\in[1,N]}$ to update position towards best position.

$$n_{i,j} = Best_j(g) \times P_{i,j} \tag{3}$$

The normalized difference $P_{i,j}$ between ith crocodile's the ith feature position and crocodile's average position. It is computed as:

$$P_{i,j} = \theta + \frac{x_{i,j} - \mu(x_i)}{Best_j(g) \times \left(UB_j - LB_j\right) + \epsilon} \tag{4}$$

where the sensitive of the exploration is controlled by the parameter θ, while ϵ maintains the floor value.

$$\mu(x_i) = \frac{1}{n}\sum_{j=1}^{n} x_{i,j} \tag{5}$$

$$R_{i,j} = \frac{Best_j(g) - x_{(rand_{\in[1,N]},j)}}{Best_j(g) + \epsilon} \tag{6}$$

$$ES(g) = 2 \times rand_{\in\{-1,1\}} \times \left(1 - \frac{1}{T}\right) \tag{7}$$

where the value 2 acts as a multiplier to provide correlation values in the range of [0, 2], and $rand_{\in\{-1,1\}}$ is a random integer between [−1, 1].

The search space is completely exploited by implementing hunting coordination and cooperation of crocodiles. It can be formulated as:

$$x_{i,j}(g+1) = \begin{cases} rand_{\in[-1,1]} \cdot Best_j(g) \cdot P_{i,j}(g), & g \le \frac{3T}{4} \text{ and } g > \frac{2T}{4} \\ [\epsilon \cdot Best_j(g) \cdot n_{i,j}(g)] - [rand_{\in[-1,1]} \cdot R_{i,j}(g)], & g \le T \text{ and } g > \frac{3T}{4} \end{cases} \quad (8)$$

2.2 Datasets

Five openly available datasets commonly utilized for intrusion detection (ID) assessment are chosen to evaluate the efficiency of MH algorithms. These datasets, widely acknowledged in the ID community [(Z. Elgamal,2022), (S. Ekinci and D. Izci,2022)], encompass Knowledge Discovery and Data Mining Cup 1999 (KDD-CUP99) (M. Tavallaee,2009), Network Security Laboratory KDD (NSL-KDD) (S. Sapre,2019), University of New South Wales - Network-Based 15 (UNSW-NB15) (A. Shiravi,2012), Canadian Institute for Cybersecurity - Intrusion Detection Evaluation Dataset 2017 (CIC-IDS2017) and CIC-IDS2018 (I. Sharafaldin,2018).

Table 1 provides a detailed overview of these datasets, including their source papers, feature counts, and sample sizes. These characteristics are essential for understanding the dataset's complexity and scale, which are critical factors in assessing intrusion detection techniques. Due to the computational demands of iterative FS methods such as MH, FS is evaluated using 10% examples of each ID dataset. Importantly, this subsampling retains the original balance between normal activities and network attacks, ensuring a representative assessment of MH algorithms' performance.

Table 1: Characteristics of intrusion detection datasets.

Dataset	Source	No. of features	No. of samples
KDD-CUP99	(M. Tavallaee,2009)	43	494,020
NSL-KDD	(S. Sapre,2019)	43	125,973
UNSW-NB15	(A. Shiravi,2012)	49	540,044
CIC-IDS2017	(I. Sharafaldin,2018)	78	2,827,876
CIC-IDS2018	(I. Sharafaldin,2018)	80	1,048,575

3 EVALUATION METRICS

Intrusion detection using reduced features generated by MH algorithms can be trained by ML models. These models can be assessed using various evaluation measures are used to determine how well an ID system is performing. Some commonly used evaluation measures in the context of ID are as follows:

True Positive (TP): It represents the number of instances where the ID system correctly identified an intrusion or attack.

True Negative (TN): It represents the number of instances where the system correctly identified non-intrusive or normal network behaviour.

False Positive (FP): It occurs when the system incorrectly flags normal network behaviour as an intrusion or attack.

False Negative (FN): It occurs when the system fails to detect an actual intrusion or attack, labelling it as normal behaviour.

Using these basic metrics, you can calculate the following evaluation measures:

Accuracy (AC): Accuracy measures of the overall correctness of the ID system and is calculated as follows:

$$AC = \frac{TP + TN}{TP + TN + FN + FP} \quad (9)$$

ID system, achieving a balance between P and R is crucial. A high P indicates that when the system flags an event as an intrusion, it is highly likely to be accurate. A high R indicates that the system is effective at detecting most of the actual intrusions. Depending on the specific requirements and priorities of the ID system, different evaluation measures may be emphasized.

4 EXPERIMENTAL RESULTS

Evaluation of RSA's ability for identifying OFS is conducted using five intrusion detection datasets, comparing its performance with other MH algorithms, including PSO (Kennedy,1995), GWO (Mirjalili,2014), MVO (Mirjalili,2016), and ROA (Jia,2021).

4.1 Experimental Setup

For this study, we implemented all the methods in Python and executed them on a computer with an Intel i7 10th generation processor, 32 GB of RAM, and running the Windows 10 system. The parameter

configurations for MH algorithms are outlined in Table 2. These settings are used based on their original research papers.

Table 2: Parameter settings for different MH algorithms.

Method	Parameters
Common settings	Population size= 32, number of runs=20, & number of iterations=100
PSO	$c_1 = c_2 = 2$, $w_{min} = 0.1$ and $w_{max} = 0.9$
GWO	C = random in [0,2], α & A decrease linearly in range [2, 0] & [1, -1]
MVO	$WEP_{max} = 1, WEP_{min} = 0.2$, α decreases from 2 to 0 and p = 6
ROA	ld=1 and β= 2
RSA	$\gamma = 0.9, \theta = 0.5$, UB & LB are vary based on the features in the dataset

This setup ensures a fair and consistent evaluation of the RSA's performance in comparison to other MH algorithms across the all datasets.

4.2 Results and Discussion

Using the real-world datasets provided in Table 1, the ability of RSA in selecting salient features is assessed together with that of other MH methods.

Table 3, presented the mean and standard deviation (STD) of fitness values of the RSA and other MH algorithms across the five datasets. It's evident that the RSA method was achieved the smallest average fitness in all five datasets, indicating superior optimization performance. The smallest STD values in all datasets indicated better stability than other MH algorithms. These results suggested that RSA was a competitive FS method, as it consistently produced best fitness values across all datasets, demonstrating its effectiveness in optimizing fitness in the context of FS.

Table 4, provided mean and STD of the number of features selected by MH algorithms for the five datasets. RSA selected the fewest mean OFS for four datasets, indicating its efficiency in FS. In the case of KDD-CUP99, both ROA and RSA have same number of features in OFS. RSA exhibited the lowest STD of the number of OFS for three datasets, indicating greater stability of FS. In UNSW-NB15 dataset, MVO and RSA showed same STD while ROA and RSA shared the same STD for CIC-IDS2018 dataset. Finally, PSO had the lowest STD for CIC-IDS2017 dataset. This analysis underscored RSA's effectiveness in selecting an OFS with lower variability, making it a strong contender in FS tasks across all datasets.

Table 3: Fitness for all datasets of all MH algorithms.

Dataset	Measure	Method				
		PSO	GWO	MVO	ROA	RSA
KDD-CUP99	Mean	0.0335	0.0220	0.0199	0.0154	**0.0094**
	STD	0.0096	0.0093	0.0073	0.0078	**0.0066**
NSL-KDD	Mean	0.0602	0.0746	0.0687	0.0612	**0.0593**
	STD	0.0081	0.0102	0.0092	0.0093	**0.0088**
UNSW-NB15	Mean	0.0372	0.0318	0.0354	0.0320	**0.0308**
	STD	0.0075	0.0057	0.0052	0.0071	**0.0049**
CIC-ID S2017	Mean	0.0136	0.0261	0.0250	0.0187	**0.0131**
	STD	0.0060	0.0084	0.0066	0.0090	**0.0082**
CIC-ID S2018	Mean	0.0340	0.0300	0.0402	0.0323	**0.0303**
	STD	0.0072	0.0094	0.0093	0.0091	**0.0061**

Table 4: Number of OFS for all datasets of all MH algorithms.

Dataset	**Measure**	**Method**				
		PSO	GWO	MVO	ROA	RSA
KDD-CUP99	Mean	40	35	41	**22**	**22**
	STD	5	9	6	7	**3**
NSL-KDD	Mean	38	34	39	37	**31**
	STD	4	6	5	5	**3**
UNSW-NB15	Mean	33	29	37	**23**	25
	STD	10	9	**4**	6	**4**
CIC-IDS2017	Mean	23	63	49	25	**21**
	STD	**3**	6	7	7	5
CIC-IDS2018	Mean	45	49	71	55	**43**
	STD	10	10	9	**8**	**8**

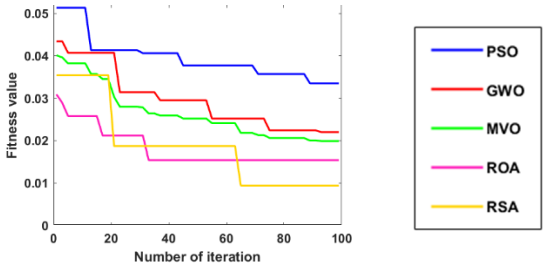
Table 5, compared mean and STD of accuracy of MH algorithms for the five datasets. The proposed RSA was outperformed the other MH methods, consistently achieving the largest average accuracy across four datasets. In terms of stability, the STD was lowest for the RSA for three datasets. This indicated that the ID systems with RSA as the FS method is highly stable and produces reliable results. In KDD-CUP99 dataset, GWO was achieved the lowest accuracy STD, followed by RSA. In summary, It highlighted RSA's effectiveness in achieving both high mean accuracy and stability across different datasets.

Table 5: Accuracy for all datasets of all MH algorithms.

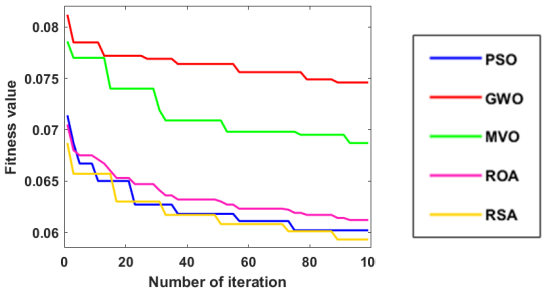
Dataset	Measure	Method				
		PSO	GWO	MVO	ROA	RSA
KDD-CUP99	Mean	0.9756	0.986	0.9895	0.9957	0.9970
	STD	0.0062	0.0061	0.0032	0.0039	0.0036
NSL-KDD	Mean	0.9481	0.9534	0.9398	0.9488	0.9528
	STD	0.0059	0.1068	0.0065	0.0068	0.0052
UNSW-NB15	Mean	0.9702	0.9747	0.9729	0.9743	0.9753
	STD	0.0051	0.0051	0.0056	0.0052	0.0134
CIC-IDS2017	Mean	0.9917	0.9884	0.9863	0.9906	0.9933
	STD	0.0058	0.0052	0.0058	0.0055	0.0050
CIC-IDS2018	Mean	0.9762	0.9812	0.9761	0.9823	0.9842
	STD	0.0173	0.0190	0.0184	0.0282	0.0072

In the comparative analysis of convergence depicted in Figure 1, after conducting 20 independent runs for each method as recommended by (Duraibi, 2023), it becomes evident that the RSA method consistently outperforms the other MH algorithms across all five datasets. The RSA method was demonstrated superior convergence rates towards optimal solutions due to its capabilities to effectively explore search space and visiting new regions in the search area , underscoring its remarkable stability and effectiveness as a FS technique for ID.

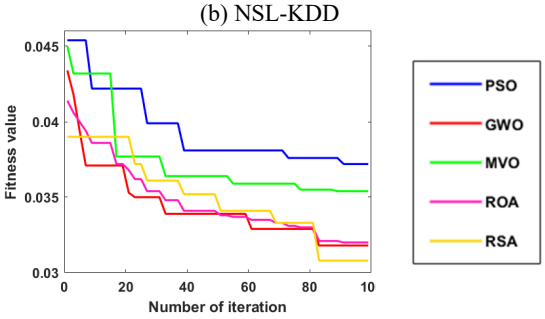
In Figure 2, we have a boxplot that displayed the performance of multiple MH algorithms across five different datasets. It visualized the distribution of accuracy across the lower, middle, and upper quartile ranges. This figure illustrated that the median accuracy achieved by the RSA algorithm surpasses that of the other MH algorithms for four datasets. In case of NSL-KDD dataset, median accuracy of GWO was slightly higher than RSA. Additionally, when considering the upper accuracy quartile, RSA outperformed the other algorithms in four out of the five datasets.



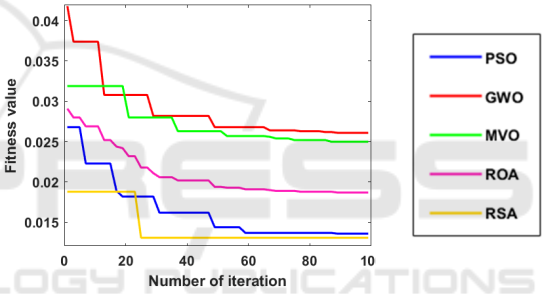
(a) KDD-CUP99



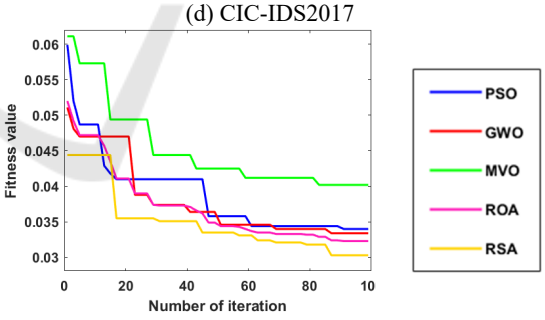
(b) NSL-KDD



(c) UNSW-NB15

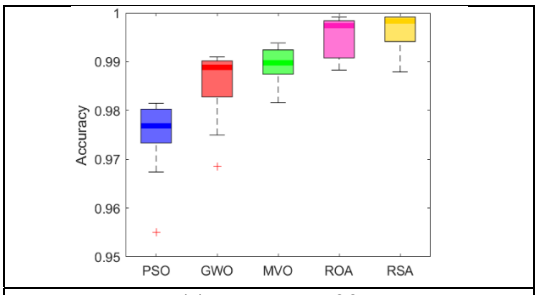


(d) CIC-IDS2017

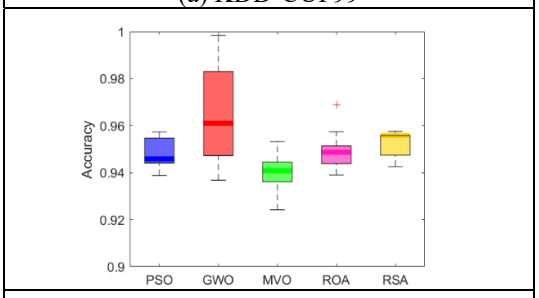


(e) CIC-IDS2018

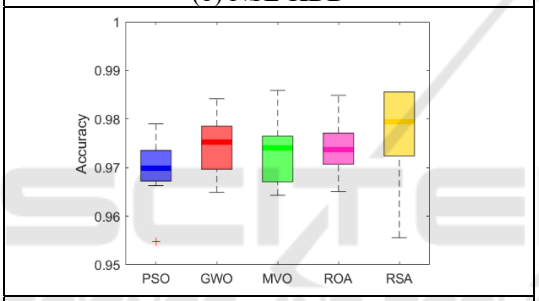
Figure 1: Convergence behaviour of all MH algorithms for ID datasets.



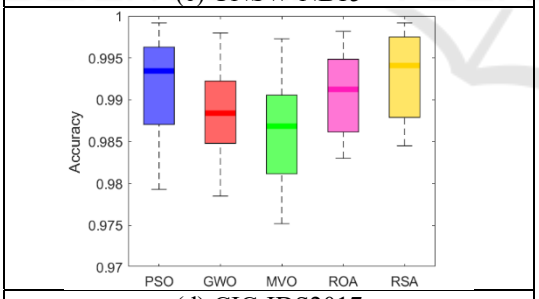
(a) KDD-CUP99



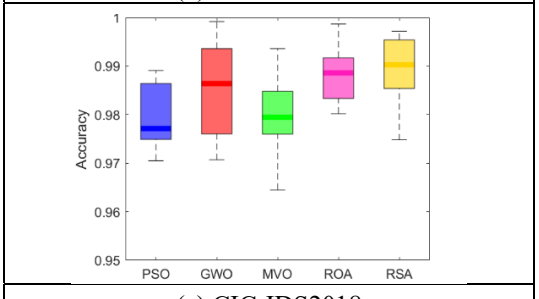
(b) NSL-KDD



(c) UNSW-NB15



(d) CIC-IDS2017



(e) CIC-IDS2018

Figure 2: Boxplots of accuracy of all MH algorithms for ID datasets.

Finding the right number of features needed for the ML task is one of the main goals of an effective FS approach. This helps to avoid selecting either too many or too few features. In a FS process, for instance, picking too many features raises the possibility of including unnecessary or redundant features, which may result in a decline in prediction accuracy. However, the RSA considered that the number of OFS in their fitness function showed better performance, and the number of the selected features were fewer Exploration of the search space and exploitation of the best solutions found are two conflicting objectives that must be taken into account when using MH algorithms. From the results provided above, RSA demonstrated a better performance in balancing exploration and exploitation factors with better convergence speed as well.

5 CONCLUSION AND FUTURE WORKS

This study presented a robust and efficient FS method using RSA for enhancing the performance of security solutions, particularly in the domain of ID. RSA offered a gradient-free approach for ID and the capabilities to avoid in getting trapped in local optima. RSA's efficacy was thoroughly examined and validated using five freely available ID datasets in the ID domain. Its performance was also rigorously compared against four other MH algorithms, including PSO, GWO, MVO, and ROA. The results demonstrated RSA's superiority, as it outperformed the other MH methods across various evaluation metrics, showcasing its capability to optimize feature subsets effectively. In future, we plan to apply RSA in other domains, including network attack detection and IoT security. Also for future consideration, RSA can be combined with another MH methods to boost its capability and produce a novel hybrid approach for solving complex identifications and classifications in ID. These developments may pave the way for the RSA to become a cornerstone method in security and optimization research.

REFERENCES

Yadav, A., Noori, M. T., Biswas, A., & Min, B. (2023). *A concise review on the recent developments in the internet of things (IoT)-based smart aquaculture practices. Reviews in Fisheries Science & Aquaculture*, 31(1), 103-118.

Khan, S. K., Shiwakoti, N., Stasinopoulos, P., & Chen, Y. (2020). *Cyber-attacks in the next-generation cars, mitigation techniques, anticipated readiness and future directions. Accident Analysis & Prevention*, 148, 105837.

Alsoufi, M. A., Razak, S., Siraj, M. M., Nafea, I., Ghaleb, F. A., Saeed, F., & Nasser, M. (2021). *Anomaly-based intrusion detection systems in iot using deep learning: A systematic literature review. Applied sciences*, 11(18), 8383.

Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). *Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecurity*, 2(1), 1-22.

Liu, H., & Lang, B. (2019). *Machine learning and deep learning methods for intrusion detection systems: A survey. applied sciences*, 9(20), 4396.

Al-Khateeb, M. O., Hassan, M. A., Al-Shourbaji, I., & Aliero, M. S. (2021). *Intelligent Data Analysis approaches for Knowledge Discovery: Survey and challenges*. Ilkogretim Online, 20(5).

Al-Shourbaji, I., Kachare, P., Fadlelseed, S., Jabbari, A., Hussien, A. G., Al-Saqqar, F., ... & Alameen, A. (2023). *Artificial Ecosystem-Based Optimization with Dwarf Mongoose Optimization for Feature Selection and Global Optimization Problems. International Journal of Computational Intelligence Systems*, 16(1), 1-24.

Fong, S., Wang, X., Xu, Q., Wong, R., Fiaidhi, J., & Mohammed, S. (2016). *Recent advances in metaheuristic algorithms: Does the Makara dragon exist?*. The Journal of Supercomputing, 72, 3764-3786.

Xu, J., & Zhang, J. (2014, July). *Exploration-exploitation tradeoffs in metaheuristics: Survey and analysis*. In Proceedings of the 33rd Chinese control conference (pp. 8633-8638). IEEE.

Kennedy, J., & Eberhart, R. (1995, November). *Particle swarm optimization*. In Proceedings of ICNN'95-international conference on neural networks (Vol. 4, pp. 1942-1948). IEEE.

Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). *Grey wolf optimizer*. Advances in engineering software, 69, 46-61.

Mirjalili, S., Mirjalili, S. M., & Hatamlou, A. (2016). *Multi-verse optimizer: a nature-inspired algorithm for global optimization*. Neural Computing and Applications, 27, 495-513.

Jia, H., Peng, X., & Lang, C. (2021). *Remora optimization algorithm. Expert Systems with Applications*, 185, 115665.

Holland, J. H. (1992). *Genetic algorithms. Scientific american*, 267(1), 66-73.

Abualigah, L., Abd Elaziz, M., Sumari, P., Geem, Z. W., & Gandomi, A. H. (2022). *Reptile Search Algorithm (RSA): A nature-inspired meta-heuristic optimizer*. Expert Systems with Applications, 191, 116158.

Z. Elgamal, A. Q. M. Sabri, M. Tubishat, D. Tbaishat, S. N. Makhadmeh et al., *Improved reptile search optimization algorithm using Chaotic map and Simulated annealing for feature selection in medical field*, IEEE Access, vol. 10, pp. 51428–51446, 2022.

S. Ekinci and D. Izci, *Enhanced reptile search algorithm with Lévy flight for vehicle cruise control system design,* Evolutionary Intelligence, vol. 2022, pp. 1–13, 2022.

M. Tavallaee, E. Bagheri, W. Lu and A. A. Ghorbani, *A detailed analysis of the KDD CUP 99 data set,* in proceedings of IEEE conference on symposium on Computational. Intelligence for Security and. Defense, Ottawa, ON, Canada, pp. 1–6, 2009.

S. Sapre, P. Ahmadi and K. Islam, *A robust comparison of the KDDCup99 and NSL-KDD IoT network intrusion detection datasets through various machine learning algorithms*, arXiv preprint, pp. 1-8, 2019.

A. Shiravi, H. Shiravi, M. Tavallaee and A. A. Ghorbani, *Toward developing a systematic approach to generate benchmark datasets for intrusion detection*, Computer. Security, vol. 31, no. 3, pp. 357–374, 2012.

I. Sharafaldin, A. H. Lashkari and A. A. Ghorbani, *Toward generating a new intrusion detection dataset and intrusion traffic characterization*, in Proceedings of the 4th International. Conference on Information Systems. Security and Privacy, Funchal, Portugal, vol. 1, no. 2, pp. 108–116, 2018.