# **Comparative analysis of whale and Harris Hawks optimization for feature selection in intrusion detection**

**Mosleh M. Abualhaj, Mohammad O. Hiari, Adeeb Alsaaidah, Mahran M. Al-Zyoud**

Department of Networks and Cybersecurity, Al-Ahliyya Amman University, Amman, Jordan



*This is an open access article under th[e CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



# *Corresponding Author:*

Mosleh M. Abualhaj Department of Networks and Cybersecurity, Faculty of Information Technology Al-Ahliyya Amman University Amman, Jordan Email: a.alsaaidah@ammanu.edu.jo

of cybersecurity.

# **1. INTRODUCTION**

The world is gradually moving into cyberspace. One key issue is the sheer number of cyberattacks launched against cyberspace [1], [2]. Several cybersecurity tools are used to protect against cyberattacks, such as the intrusion detection and prevention system (IDPS). Several techniques are adopted in the IDPS system to detect cyberattacks. One of the main techniques is signature-based, which inspects the data for a known pattern or signature of a cyberattack. Another key technique is behavior-based, which inspects the deviation of data from the usual data behavior. Besides, the IDPS systems are adopting machine learning (ML) algorithms to handle the new advanced cyberattacks, particularly those using artificial intelligence (AI) algorithms [3], [4].

ML is the part of AI that creates systems that can learn from the available data to make predictions of the incoming unrevealed data. Weaponizing the IDPS with ML improves its ability to detect cyberattacks. However, one key issue when integrating IDPS systems and ML is the large amount of data the IDPS needs to handle. As the size of data increases, the ability to detect cyberattacks is decreased [5], [6]. ML has specific types of algorithms, called feature selection algorithms, to handle large amounts of data and increase the ability of IDPS systems to detect cyberattacks. Feature selection is selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant, or noisy features. Metaheuristic algorithms are one key type of feature selection algorithms that are designed based on modeling ideas that exist in nature [7], [8]. Among the most famous metaheuristic algorithms are the whale optimization algorithm (WOA) and Harris Hawks optimization (HHO). This work will compare the performance of the HHO and WOA with the IDPS systems using the decision tree (DT) classifier [9], [10].

Several wroks have been proposed for intrusion detection. Various versions of IDPS have been developed, with many modified to enhance system efficiency by reducing attack features. This section summarizes previous studies conducted on the NSLKDD dataset with feature selection. Subba *et al.* [11] employed principal components analysis (PCA) to decrease dimensionality. The classifiers, such as support vector machines (SVM), multi-layer perceptron (MLP), C4.5, and Naïve Bayes (NB), evaluated a total of 17 features. The evaluation was conducted on both multiclass and binary-class datasets. The SVM algorithm achieved good accuracy in both multiclass and binary classification tasks. However, it should be noted that this performance was only evaluated on the particular data and may not necessarily applied to unseen data. Aziz *et al.* [12] employed the sequential floating forward selection technique to derive 26 features from the entire NSLKDD dataset. Then, the classification was applied in two layers. The first layer for normal detection uses genetic algorithm detection generation (GADG), where attacks are labeled as normal or attack. The second layer for multiclass attack classification utilizes certain classifiers such as NB, DT, J48, RF tree, and MLP, utilizing the NSLKDD and 20% KDD datasets. Each classifier exhibited optimal detection performance for a specific attack while demonstrating suboptimal performance for another type. Gupta and Kulariya [13] introduced two frameworks that utilize correlation for feature selection to identify the most pertinent feature. Additionally, they offered a feature selection method based on hypothesis testing. The performance frameworks are assessed using the KDD'99 and NSLKDD'99 datasets. Their accuracy reached 80.96% using hypothesis testing-based feature selection, with a testing time of 2.24 seconds. In their study, Lee *et al.* [14] demonstrated the significance of feature selection by examining its impact on enhancing accuracy in IDPS. The NSLKDD whole dataset is processed using an RF binary classifier in the detection model, utilizing all features without feature selection. Next, a sequential floating search is utilized to choose the optimal feature that maximizes the detection rate while minimizing false positives.

#### **2. METHOD**

### **2.1. NSLKDD dataset preprocessing**

The NSLKDD dataset will be used to assess the achievement of the WOA and HHO techniques with the help of the DT classifier. The NSLKDD dataset contains 40 features to identify the attacks [15], [16]. These features are dst\_host\_srv\_diff\_host\_rate, land, srv\_count, dst\_host\_same\_srv\_rate, num\_compromised, dst\_host\_same\_src\_port\_rate, wrong\_fragment, Count, num\_shells, same\_srv\_rate, dst\_host\_srv\_rerror\_rate, dst\_host\_count, dst\_host\_srv\_count, dst\_host\_serror\_rate, dst\_bytes, num\_failed\_logins, protocol\_type, sry\_diff\_host\_rate, service, src\_bytes, sry\_rerror\_rate, hot, logged\_in, sry\_serror\_rate, num\_outbound\_cmds, root shell, dst host rerror rate, rerror rate, su attempted, diff srv rate, num root, num file creations, num access files, dst host diff srv rate, is host login, dst host srv serror rate, urgent, is guest login, serror\_rate, Flag. Besides, the NSLKDD dataset contains 148,518 samples of attacks and benging data. The attack data is divided into four types: denial of service (DoS), probe, (R2L), and user to root (U2R). Each of the four types contains several subtypes [16]-[18]. This paper only considers the binary classification of data, classifying data as attack or benging. Therefore, all the labels in the output column will be replaced with "Attack," regardless of the attack kinds or subkinds. Thus, for binary classification, the output column will only contain attack and benign values.

Part of the NSLKDD dataset is categorical data, which hinders the function of the ML techniques. Therefore, the label-encoder algorithm replaced the categorical values with numeric values. The labelencoder algorithm replaces the values of a categorical feature with 0 to n, where n is the number of different values within the feature minus 1. For example, the "protocol\_type" feature contains three different values: tcp, udp, and icmp. The label-encoder algorithm replaces these three values with 0, 1, and 2, respectively. Another example is the label column, which contains two different values: Bengin and Attack values. The label-encoder technique replaces these two values with 0 and 1, respectively. In addition, some of the features of the NSLKDD dataset contain data distributed over wide ranges, which impacts the function of the ML techniques. Therefore, the Min-Max scaler algorithm replaces the features' data with small ranges between zero and one, which avoids the bias toward large data by ML algorithms [15], [16]. Tables 1 and 2 show a sample of the NSLKDD dataset before and after the Min-Max scaler and label-encoder techniques.

In the NSLKDD dataset preprocessing operation, the selecting feature process comes after converting data into numbers and scaling it in the same range. The HHO and WOA techniques will be used on the NSLKDD dataset to perform the feature selection process. As the name implies, the WOA simulates the whales' hunting behaviors, and the HHO simulates the hawks' hunting behaviors. The two optimizers have several advantages: i) they can be used in a wide range of optimization issues, ii) they provide robust

exploitation and exploration mechanisms, and iii) they are quite simple algorithms that can be easily understood and implemented. However, the effectiveness of these techniques can vary depending on the issue problem they are applied to. The choice between WOA and HHO, or any other optimization algorithm, depends on the characteristics of the issue at hand and the computational resources available [9], [10], [19]-[21].

The HHO and WOA techniques are widely used in cybersecurity fields, particularly with IDPS systems [9], [20]. These two techniques have been used with the NSLKDD dataset to compare their performance with IDPS. The WOA optimizer has selected 16 out of 40 features from the NSLKDD dataset: same\_srv\_rate, dst\_host\_same\_src\_port\_rate, service, srv\_diff\_host\_rate, num\_outbound\_cmds, Flag, dst\_host\_rerror\_rate, src\_bytes, is\_guest\_login, num\_failed\_logins, serror\_rate, num\_root, num\_access\_files, is host login, srv count, and srv serror rate. On the other hand, the HHO optimizer has selected 13 out of 40 features from the NSLKDD dataset: Count, num\_access\_files, dst\_host\_count, diff\_srv\_rate, dst\_host\_srv\_count, dst\_host\_same\_src\_port\_rate, dst\_host\_diff\_srv\_rate, protocol\_type, Flag, src\_bytes, dst bytes, urgent, and hot.

Table 1. Instances of the NSLKDD dataset before Min-Max scaler and label-encoder techniques

Nο	Instances	Output
1	tcp, http, SF, 300, 13788, 0	Normal
$\mathfrak{D}$	icmp, eco_i, SF, 18, 0, 0	Attack
3	tcp, http, SF, 233, 616, 0	Normal
4	tcp, http, SF, 343, 1178, 0	Normal
5	tcp, http, SF, 253, 11905, 0	Normal
6	udp, other, SF, 147, 105, 0	Normal

Tables 2. Instances of the NSLKDD dataset after Min-Max scaler and label-encoder techniques



### **2.2. Attack classification using DT**

DTs are a type of acyclic graph constructed by iteratively splitting the training data until all of the members of each partition belong to the same class label or until no more features can be partitioned. It is possible to refer to internal nodes as "Decision nodes," which denote the feature name. The leaf nodes are referred to as "Class nodes," and they are responsible for representing the correct class label. Each path of the DT describes a decision rule. Figure 1 clarifies the DT classifier [22], [23]. The DT classifier will be used to evaluate each WOA and HHO optimizer for attack classification with IDPS systems using the NSLKDD dataset.



Figure 1. Clarifies the DT classifier

*Comparative analysis of whale and Harris Hawks optimization for feature selection … (Mosleh M. Abualhaj)*

## **3. RESULTS AND DISCUSSION**

The comparison of HHO and WOA was carried out on a desktop with Ubuntu 19.10 O. S, Intel Core i9-13900KS CPU (36M Cache, up to 6.00 GHz, 24 Core), 256 GB SSD, and 16 GB RAM. Python was utilized to deploy HHO, WOA, and DT to obtain to results. The K-Fold Cross-Validation method, with k equal 5, has been used to validate the results. Five evaluation metrics, based on true positive (TP), true negative (TN), false positive (FP), and false negative (FN), were used to evaluate the WOA and HHO in enhancing the achievement of the IDPS. The metrics are accuracy (1), precision (2), recall (3), Matthews correlation coefficients (MCC) (4), and F1-Score (5) [24], [25].

$$
Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}
$$
 (1)

$$
Precision = \frac{TP}{(TP + FP)}
$$
 (2)

$$
\text{Recall} = \frac{TP}{(TP + FN)}\tag{3}
$$

$$
MCC = \frac{((TP*TN) - (FP*FN))}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}
$$
\n
$$
(4)
$$

$$
F1-score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
$$
 (5)

Figure 2 shows the accuracy of the WOA compared to the HHO. The WOA achieved an accuracy of 97.50%, while the HHO achieved an accuracy of 97.59%. Clearly, the HHO outperforms the WOA by 0.09% in terms of accuracy in detecting the attack by the IDPS systems using the NSLKDD dataset and DT classifier.

Figure 3 shows the precision of the WOA compared to the HHO. The WOA achieved a precision of 97.50%, while the HHO achieved a precision of 97.59%. Clearly, the HHO outperforms the WOA by 0.09% in terms of precision in detecting the attack by the IDPS systems using the NSLKDD dataset and DT classifier.



Figure 2. Accuracy of the WOA and HHO optimizers



Figure 4 shows the recall of the WOA compared to the HHO. The WOA achieved a recall of 97.50%, while the HHO achieved a recall of 97.59%. Clearly, the HHO outperforms the WOA by 0.09% in terms of recall in detecting the attack by the IDPS systems using the NSLKDD dataset and DT classifier.

Figure 5 shows the MCC of the WOA compared to the HHO. The WOA achieved an MCC of 96.02%, while the HHO achieved an MCC of 96.17%. Clearly, the HHO outperforms the WOA by 0.15% in terms of MCC in detecting the attack by the IDPS systems using the NSLKDD dataset and DT classifier.

Figure 6 shows the F1-Score of the WOA compared to the HHO. The WOA achieved an F1-Score of 97.50%, while the HHO achieved an F1-Score of 97.59%. Clearly, the HHO outperforms the WOA by 0.09% in terms of F1-Score in detecting the attack by the IDPS systems using the NSLKDD dataset and DT classifier.



Figure 4. Recall of the WOA and HHO optimizers Figure 5. MCC of the WOA and HHO optimizers



Figure 6. F1-Score of the WOA and HHO optimizers

In summary, the HHO optimization algorithm outperforms the WOA optimization algorithm in terms of enhancing the achievement of the IDPS systems with the five evaluation metrics. In particular, accuracy, recall, precision, and F1-Score have all gone up by 0.09% and MCC by 0.15% when the HHO optimization algorithm for feature selection is used instead of the WOA optimization algorithm in IDPS systems. The superior performance of HHO over WOA in terms of feature selection and resulting accuracy can be attributed to WOA's balanced approach to exploration and exploitation, its ability to maintain diversity, adaptive position updates, and an effective encircling mechanism. These characteristics enable WOA to select a more relevant, non-redundant, and generalizable subset of features, leading to better model

*Comparative analysis of whale and Harris Hawks optimization for feature selection … (Mosleh M. Abualhaj)*

performance and higher accuracy. Therefore, the WOA optimization algorithm and DT classifier prove to be highly efficient tools with the IDPS systems, offering superior performance and reliability.

#### **4. CONCLUSION**

In conclusion, this research has methodically investigated the use of WOA and HHO for feature selection in IDPS using the NSL-KDD dataset. Our findings highlight the commendable performance of both optimization algorithms in enhancing the accuracy of IDPS when coupled with the DT classification model. The experimental results reveal that DT with HHO outshines its WOA counterpart, achieving an accuracy of 97.59% compared to 97.5%. This disparity underscores the nuanced efficiency of optimization algorithms in the specific context of feature selection for intrusion detection. The study not only reaffirms the potential of nature-inspired algorithms in cybersecurity applications but also emphasizes the need for careful algorithm selection based on the intricacies of the problem domain. Future work could delve deeper into fine-tuning parameters, exploring other datasets, and extending the analysis to diverse ML models. Overall, the insights gleaned from this research contribute to the ongoing discourse on optimization algorithms in cybersecurity, paving the way for continued advancements in intrusion detection methodologies. Future works will compare other optimization algorithms for feature selection. In addition, more classifiers will be tested with the WOA optimization algorithm.

#### **REFERENCES**

- [1] M. Bang and H. Saraswat, "Building an effective and efficient continuous web application security program," in 2016 *International Conference On Cyber Situational Awareness, Data Analytics And Assessment (CyberSA),* 2016, pp. 1-4, doi: 10.1109/CyberSA.2016.7503287.
- [2] M. Abualhaj, M. Al-Zyoud, M. Hiari, Y. Alrabanah, M. Anbar, A. Amer, and A. Al-Allawee, "A fine-tuning of decision tree classifier for ransomware detection based on memory data," *International Journal of Data and Network Science,* vol. 8, no. 2, pp. 733-742, 2024.
- [3] P. R. Chandre, P. N. Mahalle, and G. R. Shinde, "Machine learning based novel approach for intrusion detection and prevention system: a tool based verification," in *IEEE Global Conference on Wireless Computing and Networking (GCWCN),* 2018, pp. 135- 140, doi: 10.1109/GCWCN.2018.8668618.
- [4] M. Abualhaj, A. Abu-Shareha, Q. Shambour, A. Alsaaidah, S. Al-Khatib, and M. Anbar, "Customized K-nearest neighbors' algorithm for malware detection," *International Journal of Data and Network Science,* vol. 8, no. 1, pp. 431-438, 2024.
- [5] P.-Z. Zhou, H. Zhang, and W. Liang, "Research on hybrid intrusion detection based on improved harris hawk optimization algorithm," *Connection Science,* vol. 35, no. 1, pp. 1-24, Apr. 2023, doi: 10.1080/09540091.2023.2195595.
- [6] G. Yedukondalu, G. H. Bindu, J. Pavan, G. Venkatesh, and A. SaiTeja, "Intrusion detection system framework using machine learning," in *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA),* Coimbatore, India, 2021, pp. 1224-1230, doi: 10.1109/ICIRCA51532.2021.9544717.
- [7] H. Mendes, S. E. Quincozes, and V. E. Quincozes, "A web user interface tool for metaheuristics-based feature selection assessment for IDSs," Oct. 2022, doi: 10.1109/csnet56116.2022.9955616.
- [8] M. Kolhar, F. Al-Turjman, A. Alameen, and M. M. Abualhaj, "A three layered decentralized IoT biometric architecture for city lockdown during COVID-19 outbreak," *Ieee Access,* vol. 8, pp. 163608-163617, 2020.
- [9] M. Alazab, R. A. Khurma, P. Castillo, B. Abu-Salih, A. Martín, and D. Camacho, "An effective networks intrusion detection approach based on hybrid harris hawks and multi-layer perceptron," *Egyptian Informatics Journal,* vol. 25, no. 1, pp. 1-9, Mar. 2024, doi: 10.1016/j.eij.2023.100423.
- [10] S. Mirjalili and A. Lewis, "The whale optimization algorithm," *Advances in Engineering Software*, vol. 95, pp. 51–67, May 2016, doi: 10.1016/j.advengsoft.2016.01.008.
- [11] B. Subba, S. Biswas, and S. Karmakar, "Enhancing performance of anomaly based intrusion detection systems through dimensionality reduction using principal component analysis," in *International Conference on Advanced Networks and Telecommunications Systems,* IEEE, 2017.
- [12] S. A. Aziz, A. S. EL-Ola Hanafi, and A. E. Hassanien, "Comparison of classification techniques applied for network intrusion detection and classification," *Journal of Applied Logic,* vol. 24, pp. 109–118, 2017.
- [13] P. Gupta and M. Kulariya, "A framework for fast and efficient cyber security network intrusion detection using apache spark," in *International Conference on Advances in Computing and Communications,* 2016, pp. 824–831.
- [14] J. Lee, D. Park, and C. Lee, "Feature selection algorithm for intrusions detection system using sequential forward search and random forest classifier," *Transactions On Internet And Information Systems,* vol. 11, no. 10, pp. 5132–5148, 2017.
- [15] H. Al-Mimi *et al.,* "An enhanced intrusion detection system for protecting HTTP services from attacks," *International Journal of Advances in Soft Computing and Its Applications,* vol. 15, no. 3, 2023.
- [16] Ü. Çavuşoğlu, "A new hybrid approach for intrusion detection using machine learning methods," *Applied Intelligence,* vol. 49, pp. 2735–2761, 2019, doi: 10.1007/s10489-018-01408-x.
- [17] M. M. Abualhaj *et al.,* "A paradigm for DoS attack disclosure using machine learning techniques," *International Journal of Advanced Computer Science and Applications,* vol. 13, no. 3, 2022.
- [18] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A deep learning approach to network intrusion detection," *IEEE Transactions on Emerging Topics in Computational Intelligence,* vol. 2, no. 1, pp. 41-50, Feb. 2018, doi: 10.1109/TETCI.2017.2772792.
- [19] R. A. Khurma, M. A. Awadallah, and I. Aljarah, "Binary harris hawks optimisation filter based approach for feature selection," *In 2021 Palestinian International Conference on Information and Communication Technology (PICICT),* Sep. 2021, doi: 10.1109/picict53635.2021.00022.
- [20] B. D. Shivahare *et al.,* "Survey paper: whale optimization algorithm and its variant applications," *In 2021 International Conference on Innovative Practices in Technology and Management (ICIPTM)*, Feb. 2021, doi: 10.1109/iciptm52218.2021.9388344.
- [21] M. Nadimi-Shahraki *et al.,* "A systematic review of the whale optimization algorithm: theoretical foundation, improvements, and hybridizations," *Archives of Computational Methods in Engineering,* vol. 30, pp. 4113–4159, 2023, doi: 10.1007/s11831-023- 09928-7
- [22] M. M. Abualhaj and S. N. Al-Khatib, "Using decision tree classifier to detect trojan horse based on memory data," TELKOMNIKA, vol. 22, no. 2, pp. 393–393, Apr. 2024, doi: [10.12928/telkomnika.v22i2.25753.](https://doi.org/10.12928/telkomnika.v22i2.25753)
- [23] X.-Y. Shih, Y. Chiu, and H. -E. Wu, "Design and implementation of decision-tree (DT) online training hardware using dividerfree GI calculation and speeding-up double-root classifier," in *IEEE Transactions on Circuits and Systems I: Regular Papers,* vol. 70, no. 2, pp. 759-771, Feb. 2023, doi: 10.1109/TCSI.2022.3222515.
- [24] M. Gharib, T. Zoppi, and A. Bondavalli, "On the properness of incorporating binary classification machine learning algorithms into safety-critical systems," in *IEEE Transactions on Emerging Topics in Computing*, vol. 10, no. 4, pp. 1671-1686, 1 Oct.-Dec. 2022, doi: 10.1109/TETC.2022.3178631.
- [25] H. E. Massari, Z. Sabouri, S. Mhammedi, and N. Gherabi, "Diabetes prediction using machine learning algorithms and ontology," *in Journal of ICT Standardization,* vol. 10, no. 2, pp. 319-337, 2022, doi: 10.13052/jicts2245-800X.10212.

## **BIOGRAPHIES OF AUTHORS**



**Prof. Mosleh M. Abualhaj <b>in Al-Ahliyya C** is a senior lecturer in Al-Ahliyya Amman [University.](https://www.google.jo/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CBwQFjAA&url=http%3A%2F%2Fwww.ammanu.edu.jo%2F&ei=6BrTVOeqOdbfaorvgIgP&usg=AFQjCNHrdvgqrBR3Xl1DYmzuaEkGuW9Mwg&bvm=bv.85464276,bs.1,d.bGQ&cad=rja) He received his first degree in Computer Science from Philadelphia University, Jordan, in 2004, master degree in Computer Information System from the Arab Academy for Banking and Financial Sciences, Jordan in 2007, and Ph.D. in Multimedia Networks Protocols from Universiti Sains Malaysia in 2011. His research area of interest includes VoIP, congestion control, and cybersecurity data mining and optimization. He can be contacted at email: m.abualhaj@ammanu.edu.jo.



**Mr. Mohammad O. Hiari <b>in is a** lecturer in [Al-Ahliyya Amman University.](https://www.google.jo/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CBwQFjAA&url=http%3A%2F%2Fwww.ammanu.edu.jo%2F&ei=6BrTVOeqOdbfaorvgIgP&usg=AFQjCNHrdvgqrBR3Xl1DYmzuaEkGuW9Mwg&bvm=bv.85464276,bs.1,d.bGQ&cad=rja) He received his first degree in Software Engineering from Philadelphia University, Jordan, in August 2004 and master degree in Computer Science from Al Balqa Applied University, Jordan in February 2016. His research area of interest includes VoIP and cybersecurity data mining and optimization. He can be contacted at email: m.hyari@ammanu.edu.jo.



**Dr. Adeeb Alsaaidah <b>republik republik** received the bachelor's degree in Computer Engineering from the Faculty of Engineering, ALBalqa Applied University, the master's degree in Networking and Computer Security from NYIT University, and the Ph.D. degree in Computer Network from USIM, Malaysia. He is currently an Assistant Professor in Network and Cybersecurity department at Al-Ahliyya Amman University (AAU). His research interests include network performance, multimedia networks, network quality of service (QoS), the IoT, network modeling and simulation, network security, and cloud security. He can be contacted at email: a.alsaaidah@ammanu.edu.jo.



**Dr. Mahran M. Al-Zyoud <b>is a C** is an assistant professor of Networks and Cybersecurity at Al-Ahliyya Amman University. He received a B.Sc. in Computer Science and a M.Sc. in Computer Information Systems from the University of Jordan in 2004 and 2012. He received a Ph.D. in Computer Science from the University of Alabama, USA, in 2019. His research interests include data privacy and IoT security. He can be contacted at email: [m.zyoud@ammanu.edu.jo.](mailto:m.zyoud@ammanu.edu.jo)

*Comparative analysis of whale and Harris Hawks optimization for feature selection … (Mosleh M. Abualhaj)*