# A Salp Swarm Algorithm for Interpreting Model Predictions

*Alia* A. Hussein[1], *Ali* J. Ramadhan[1,2*], *Ali* TaeiZadeh[2] and *Mohand* Hussein Issa[3]

[1]University of Alkafeel, Najaf, Iraq
[2] University of Qom, Qom, Iran
[3]Ministry of Education, Directorate of Education, Najaf, Iraq

**Abstract.** The Internet of Things (IoT), is changing practically every aspect of modern life. The proliferation of IoT has led to a rise in the frequency of cyber catastrophes. The threat landscape that security professionals face is dynamic, complex, and diversified. This paper proposes a novel approach to enhance Internet of Things applications by fusing the swarm intelligence of Salp Swarm Algorithms (SSA) with the predictive power of Random Forest (RF) and Decision Tree (DT) models Even though there is a lot of interest in the topic of explainable Artificial Intelligence (XAI) these days, more research is still needed to fully understand how successful XAI is at finding attack surfaces and vectors when implemented in cyber security applications. The growing use of machine/deep learning models in cyber defense, especially anomaly-based IDS, requires understanding the architecture of the models and providing evidence for their predictions to determine the probability of intrusions. Numerous approaches to interpretation have been proposed. They help researchers comprehend things like which variables have influenced the machine learning predictions. In this paper, we primarily address two popular local interpretation methods in machine learning: Shapley values and Local Interpretable Model-Agnostic Explanations (LIME).

## 1 Introduction

The term "Internet of Things," or "IoT," refers to a group of gadgets that have been updated with CPUs and network cards to facilitate data interchange and Internet access. The term "things" here refers to a broad category of objects that can be controlled by users through the internet, applications, or other interfaces, ranging from sophisticated cloud servers to basic sensors [1]. Large data processing techniques are needed for the efficient processing of data from those networks because of the widespread use of different IoT devices and the subsequent generation of multimodal and high-dimensional data. One paradigm that is used to help with these problems is artificial intelligence (AI). Artificial Intelligence (AI) technologies, including machine learning (ML) and deep learning (DL), are extensively employed across various sectors and yield favorable outcomes when processing vast volumes of data.

* Corresponding author: ali.j.r@alkafeel.edu.iq

Because of their unique properties, IoT networks require different protection tactics than those used in traditional business environments. Several security protocols have been put forth to guard IoT networks and stop criminal activity. Traditional techniques include user authentication, data encryption, firewalls, and anti-malware software. Moreover, since more and more user types gain from knowing the underlying causes of intrusion detection, an explanation of the predictions made by ML-based intrusion detection systems is necessary. The field of explainable artificial intelligence (XAI) emerged in response to the increasing demand for the explanation of machine learning models for application in cyber protection. XAI sheds light on the black-box concept by making its workings and forecasts more understandable. One of the main shortcomings of the existing DL-based intrusion detection systems is the identification of zero-day attacks and the interpretation of how the models were able to detect them [4][5]. Researchers have focused on understanding the models to comprehend the judgments made by machine learning models [6].

## 2. PROPOSED METHODOLOGY

This section displays the suggested architecture for an explainable AI-based IDS platform. Finding a remedy for malicious conduct in IDS requires knowledge of this type of decision, which is only the first stage in the process. Administrators can identify the network segment, feature, and security policy that has been compromised by an attacker by having a deeper understanding of the decision taken. With the aid of the data provided by XAI, the IDS operator can take the necessary actions, such as troubleshooting the IDS model or putting new security measures in place to prevent similar assaults in the future. The proposed framework aims to provide an effective explainable AI-based intrusion detection system (IDS) that can generate both local and global explanations by utilizing multiple XAI techniques. The local approach focuses on explaining a specific network transaction that was found to be malicious (refer to Figure 1).

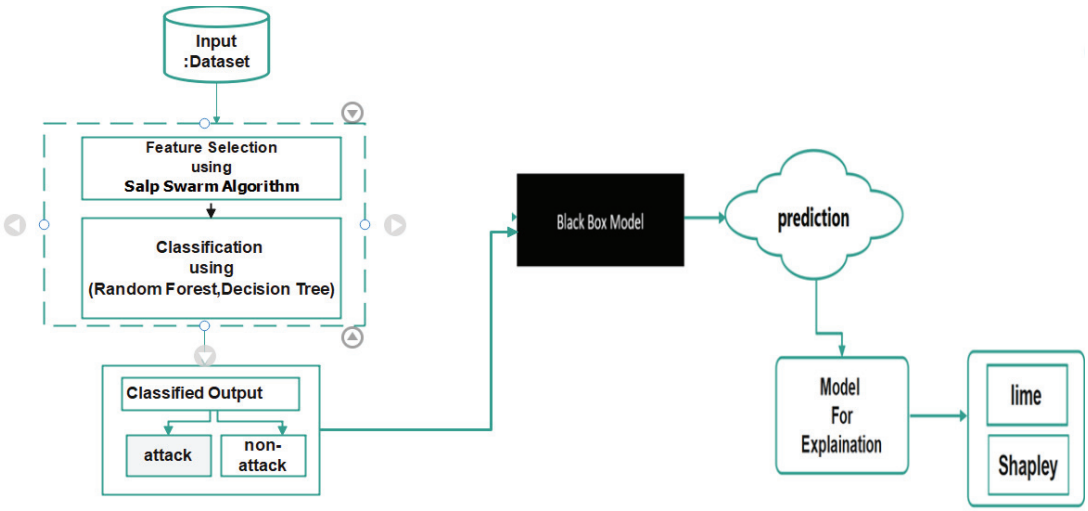


**Fig. 1.** Propose System

## 3. FEATURE SELECTION

We introduce the Salp Swarm Algorithm (SSA) of Mirjalili et al. [7]. The SSA belongs to the class of metaheuristic swarm-based algorithms. The movement patterns of sea salp population chains are imitated by the SSA, a swarm intelligence optimization system [8]. The primary goal of optimization techniques is to maximize the objective function or fitness function to identify the optimal choice (problem-solution). Finding the best value among several feasible options is the aim of the decision-making process. The best option or value among all the possibilities is chosen as a result of the optimization procedure [9]. Figures 2

and 3 [9] illustrate individual salp animals, salp chains, and the ideas of leader and follower. Table 1 presents the pseudocode for the salp swarm algorithm (SSA) [9].

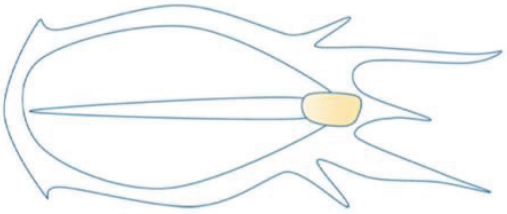

**Fig .2.** An example of a single salp animal [9].

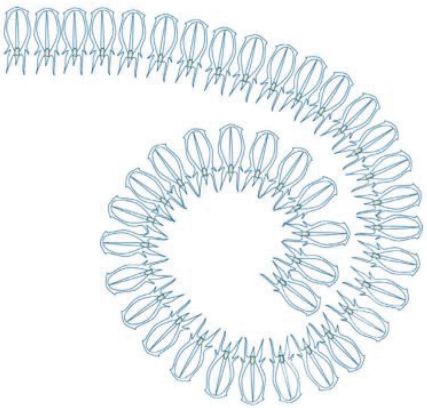

**Fig .3.** Illustration of a salp chain with the idea of a leader and a follower [9].

**Table 1.** Algorithm of the pseudocode for the salp swarm algorithm (SSA) [9]

| No. | Input parameter: Population Size, Number of Iterations, Min Values, Max Values |
|---|---|
| 1 | Initialize the salp population X_i (i = 1, 2, 3, . . . , n) considering ub and lb |
| 2 | while (end condition is not satisfied) do |
| 3 | calculate the fitness of each search agent (salp) |
| 4 | F = the best search agent |
| 5 | update c1 by Equation (3) |
| 6 | for each salp (X_i) |
| 7 | if (i == 1) |
| 8 | Revise the leading salp's position by Equation (2) |
| 9 | Else |
| 10 | Update the position of the follower salp by Equation (5) |
| 11 | End |
| 12 | End |
| 13 | Amend the salps based on the upper and lower bounds of variables |
| 14 | End |
| 15 | Return f |
| Output: Global best solution | |

## 3. PERFORMANCE METRICS

This study includes all impacts in the accuracy, or F1-Score, PPR, DR, and MCC, because it consistently integrates correctness and recalls into a single value for evaluating the entire system's performance, see Table 2. The number of attack episodes that are classified as attacks is known as True Positives (TP). The number of normal occurrences that are correctly categorized as normal is known as True Negatives (TN). False Positives (FP) are the number of ordinary occurrences that are mistakenly classified as attacks. The number of real assault cases mistakenly classified as routine occurrences is known as False Negatives (FN).

**Table 2.** Confusion Matrix

| Predicted | Actual positive | Actual negative |
|---|---|---|
| Positive | TP | FP |
| Negative | FN | TN |

The accuracy value can be calculated directly from the confusion matrix in Equation (1.1) [15].

$$\text{Accuracy (ACC)} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \times 100\% \quad (1)$$

This criterion is assessed using Equation (1.2) [16]. The division of all samples predicts true positive labels on all samples with a true positive and false negative.

$$\text{Precision (PPR)} = \frac{Tp}{FP+Tp} \times 100\ \% \quad (2)$$

$$\text{Detection Rate (DR)} = \frac{Tp}{TP+FN} \times 100\% \quad (3)$$

The F1-score is crucial for comparing performance between classes because it is calculated independently for each class resolution and accuracy. It can be calculated using Equation (1.4) after obtaining precision and recall [17].

$$\text{Recall} = \frac{\text{TP}}{(\text{TP+FN})} \times 100\% \quad (4)$$

$$\text{F1-Score} = 2*\left(\frac{\text{PPR}*\text{SEN}}{\text{PPR+SEN}}\right)\times 100\% \quad (5)$$

MCC uses the confusion matrix's TP, TN, FP, and FN factors. Equation (2.7) can be used to calculate the MCC score.

$$\text{MCC} = \frac{(\text{TP}*\text{TN} - \text{FP}*\text{FN})}{\sqrt{(\text{TP+FP})*(\text{TP+FN})*(\text{TN+FP})*(\text{TN+FN})}} \times\ 100\% \quad (6)$$

The values produced by the MCC score range from -1 to 1. MCC = 1 represents a classification that is 100% correct, while MCC = -1 represents a classification that is 100% erroneous.

## 4. RESULTS AND DISCUSSIONS

The results in Table 3 demonstrate how well the SSA algorithm fared in terms of DR, F1, and Accuracy in the IoT-MQTT dataset. In terms of outcomes, SSA+RF performed better than SSA+DT.

**Table 3**. Analysis of SSA+RF and SSA+DT method

| Dataset | Method | ACC | F1 | DR | MCC |
|---|---|---|---|---|---|
| Biflow | SSA+DT | 95.54 | 91.92 | 90.70 | 88.86 |
| | SSA+RF | 96.19 | 93.03 | 93.29 | 90.41 |

Figure 4 demonstrates that the SSA-DT model has classified a total of 16383 cases into class 0 and 5921 instances into class 1.

In Figure 4 the Shapley value for each feature means how much it contributes to the prediction compared to other features the "Var5" have the highest Shapley values, indicating that they are the most important features for predicting." Var4" have the low Shapley value indicates that their presence has relatively little effect of prediction.

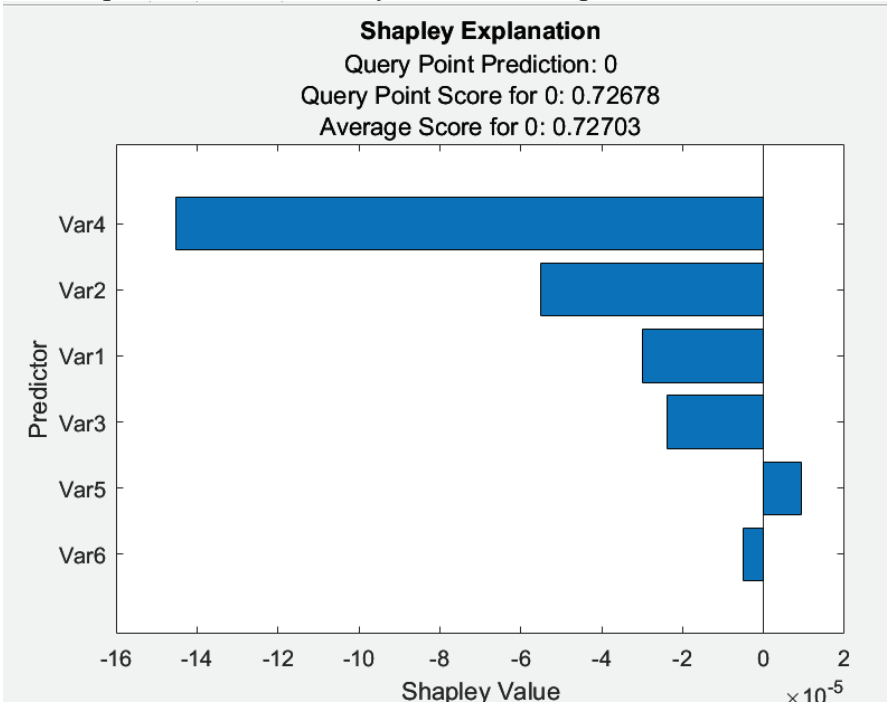


**Fig. 4.** The Shapley values for the predicted class of the IoT-MQTT dataset.

Table 4 represents the impact of each feature on the prediction, positive coefficient that means an increase in that variable is linked to an increase in the prediction, negative coefficient that means an increase in that variable is linked to a reduction in the forecast.

**Table 4.** The Shapley Values all features for each class of IoT-MQTT

| Predictor | 1 | 0 |
|---|---|---|
| Var1 | -0.0047 | 0.0047 |
| Var2 | -0.04081 | 0.0408 |
| Var3 | -0.01808 | -0.0148 |
| Var4 | -0.045901 | 0.045901 |
| Var5 | 0.09652 | -0.09652 |
| Var6 | -0.04165 | 0.054784 |

Figure 5 represents the importance of the features among the 95 features resulting from the purpose model. The "Var2" feature is the most effective, contributing the most to malware identification via classification. However, the " Var1" feature is the least efficient and can deliver the worst results for the suggested method.

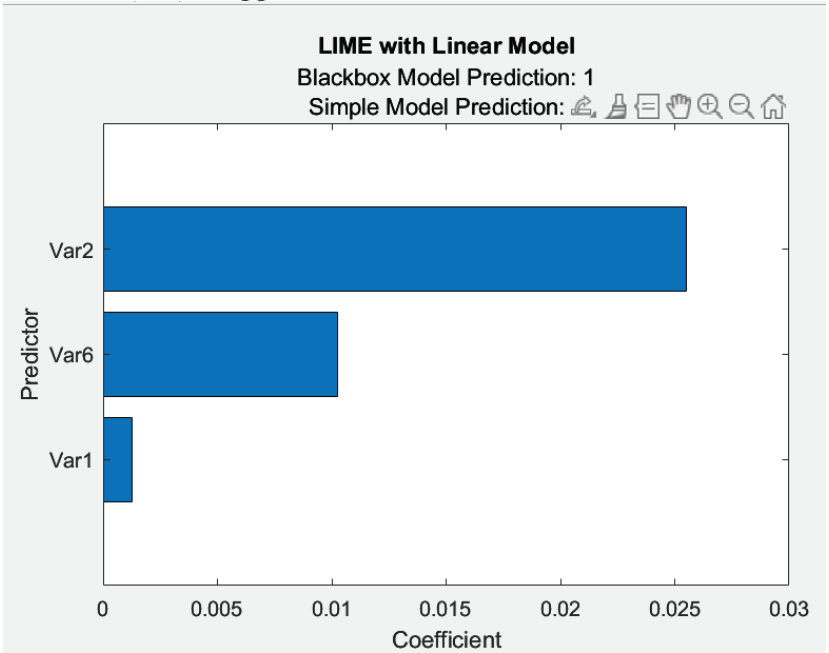


**Fig.5.** Most significant features of the IoT-MQTT dataset.

## 5. CONCLUSION

An effective machine learning technique for intelligent intrusion detection in the Internet of Things (IoT) is swarm-based feature selection. Ant colony optimization and particle swarm optimization are two instances of swarm intelligence approaches that are used to address the challenges brought on by the massive dimensionality and dynamic nature of IoT data. Swarm-based feature selection has several benefits, including finding relevant features, lowering dimensionality, and quickly scanning a vast feature space. This improves the intrusion detection system's interpretability as well as the model's performance. The optimal qualities must be selected to build dependable and effective models that can accurately identify aberrant activity in IoT networks. Swarm-based methodology also meshes well with Internet of Things settings' flexible and self-organizing features. Because these algorithms can adapt to changing threat scenarios and network conditions, they are perfect for dynamic, real-time intrusion detection applications. The efficiency of intrusion detection systems in the Internet of Things could be greatly increased by combining machine learning with swarm-based feature selection. An interesting direction for interpretable machine learning is the potential combination of the Salp Swarm Algorithm with well-known interpretability methods like SHAP and LIME. This strategy opens the door to a more thorough comprehension of model predictions by fusing the advantages of swarm intelligence with the accuracy and lucidity offered by traditional techniques. It is important that scholars and professionals persist in investigating and verifying these comprehensive methods for analyzing intricate machine learning frameworks.

## REFERENCES

1. M. A. Al-Garadi, A. Mohamed, A. K. Al-Ali, X. Du, I. Ali, and M. Guizani, “A survey of machine and deep learning methods for internet of things (IoT) security,” IEEE Commun. Surv. Tutorials, vol. **22**, no. 3, pp. 1646–1685, 2020.
2. I. H. Sarker, M. H. Furhad, and R. Nowrozy, “Ai-driven cybersecurity: an overview,

security intelligence modeling and research directions," SN Comput. Sci., vol. **2**, pp. 1–18, 2021.
3. M. M. Hassan, A. Gumaei, A. Alsanad, M. Alrubaian, and G. Fortino, "A hybrid deep learning model for efficient intrusion detection in big data environment," Inf. Sci. (Ny)., vol. **513**, pp. 386–396, 2020.
4. S. Mohammadi, H. Mirvaziri, M. Ghazizadeh-Ahsaee, and H. Karimipour, "Cyber intrusion detection by combined feature selection algorithm," J. Inf. Secur. Appl., vol. **44**, pp. 80–88, 2019.
5. N. Moustafa, "A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets," Sustain. Cities Soc., vol. **72**, p. 102994, 2021.
6. C. Wu, A. Qian, X. Dong, and Y. Zhang, "Feature-oriented design of visual analytics system for interpretable deep learning based intrusion detection," in 2020 International Symposium on Theoretical Aspects of Software Engineering (TASE), 2020, pp. 73–80.
7. S. Mirjalili, A. H. Gandomi, S. Z. Mirjalili, S. Saremi, H. Faris, and S. M. Mirjalili, "Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems," Adv. Eng. Softw., vol. **114**, pp. 163–191, 2017.
8. J. Zhang and J.-S. Wang, "Improved salp swarm algorithm based on levy flight and sine cosine operator," Ieee Access, vol. **8**, pp. 99740–99771, 2020.
9. L. Abualigah, M. Shehab, M. Alshinwan, and H. Alabool, "Salp swarm algorithm: a comprehensive survey," Neural Comput. Appl., vol. **32**, pp. 11195–11215, 2020.
10. M. T. Ribeiro, S. Singh, and C. Guestrin, "' Why should i trust you?' Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
11. S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," May 2017, (accessed on 19 March 2024). [Online]. Available: http://arxiv.org/abs/1705.07874
12. D. B. Gillies, Some theorems on n-person games. Princeton University, 1953.
13. A. Joseph, "Shapley regressions: A framework for statistical inference on machine learning models," 2019.
14. T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction, vol. 2. Springer, 2009.
15. H. Hindy, E. Bayne, M. Bures, R. Atkinson, C. Tachtatzis, and X. Bellekens, "Machine learning based IoT intrusion detection system: An MQTT case study (MQTT-IoT-IDS2020 dataset)," in International networking conference, 2020, pp. 73–84.
16. H. Hindy, C. Tachtatzis, R. Atkinson, E. Bayne, and X. Bellekens, "Mqtt internet of things intrusion detection dataset." Jun, 2020.
17. M. Buckland and F. Gey, "The relationship between recall and precision," J. Am. Soc. Inf. Sci., vol. **45**, no. 1, pp. 12–19, 1994.