FEATURE SELECTION USING BIO-INSPIRED OPTIMIZATION FOR IOT INTRUSION DETECTION AND PREVENTION SYSTEM

*Richa Singh *, R. L. Ujjwal*

USIC&T, G.G.S.I.P.U, New Delhi
India

* Corresponding Author, e-mail: richa.singh081991@gmail.com

Abstract: Nowadays smart Internet of Things (IoT) devices are used briskly, and these devices communicate with each other via wireless medium. However, this increase in IoT devices has resulted in a rise of security issues associated with the IoT system. Therefore, an intrusion detection and prevention system (IDPS) is used to locate and report any malicious activity. The IDPS's feature selection (FS) task is necessary to improve the data quality and decrease the data used for classifying intrusive traffic. Therefore, this paper proposes a novel FS method that hybridizes improved salp swarm algorithm and harris hawk optimization algorithm. The XGBoost classifier is used for classifying reduced network traffic. Proposed system demonstrates high accuracy and low computation time, surpassing other related approaches used for the IDPS feature selection task.

Key words: security, intrusion detection and prevention system, IoT, improved salp swarm algorithm, harris hawk optimization.

1. INTRODUCTION

The modern world is highly dependent on the technology advancement due to the terminology “Internet of things” and achievements in the field of artificial intelligence (AI) [1]. An IoT system consists of many smart devices equipped with sensors, processors, and hardware for communication. These devices can perceive, send and take action on data obtained from these smart devices. The number of IoT devices is increasing rapidly which simulates the security issues involved with the IoT devices. Therefore, the security of IoT devices is crucial, given the increasing number of devices and potential threats to their security. To address these security issues, various solutions have been proposed, including access control, intrusion detection and prevention system (IDPS), authentication, and firewalls. IDPS involves identifying any unknown activity that affects the privacy and confidentiality of the system and reporting them. The data obtained from sensing devices is immense and diversified. Feature selection (FS) is a vital process in data analysis, that selects the essential features while ignoring the rest. Bio-inspired algorithms [2], which are inspired by natural processes such as genetic

algorithms or swarm intelligence, have gained popularity in FS because they can efficiently search for optimal feature subsets.

Therefore, the main contribution of this paper includes the proposal of a novel hybridized FS approach that utilizes the improved salp swarm algorithm (ISSA) and the harris hawk optimization (HHO) algorithm. The reduced optimal network traffic is provided as an input to the task of intrusion detection and classification using the XGBoost classifier. Here, multiclass classification is performed. Subsequently, the prevention of invasive traffic is done. The proposed system's effectiveness is determined against two datasets, BoT-IoT and IoTID20.

Remaining paper is described as follows: Section 2 describes the literature survey. Section 3 describes the materials and methods used for this research. Proposed system is presented in Section 4 while Section 5 covers implementation and result part. Finally, Section 6 presents the conclusion.

2. LITERATURE SURVEY

This section discusses recent research on using bio-inspired algorithms for implementing intrusion detection systems (IDS). The work in [3], proposed the FS using the bird swarm algorithm(BSA) and gorilla troops optimizer (GTO). However, the hybridized GTO-BSA takes more computation time in detecting optimal features compared to other approaches. Whale optimization (WO) algorithm, with modified using transfer functions is proposed in [4], for the FS of IoT-based IDS. The proposed system's performance is better than other similar approaches. However, the proposed work does not outperform for all datasets in term of specificity, and false positive. The work in [5] proposed an anomaly detection system based on ensemble learning. The optimal FS from the network traffic is performed using the PSO algorithm. Afterward, the hybrid ensemble method is used for the classification of reduced data using gradient boosting machine (GBM), and bootstrap aggregation. Furthermore, FS done by reptile search algorithm (RSA) for the IoT-based IDS is proposed by authors [6]. In this paper, first, the features are extracted using deep learning-based CNN. Afterward, RSA is used for selecting essential features. The performance of a system is evaluated against multiple datasets. However, the RSA has slow convergence speed. Another work in [7] uses a binary form of farmland fertility algorithm for essential FS. Here, a V-shaped function is employed for converting the continuous space to a binary one. The proposed system is evaluated against UNSNB15 and NSL-KDD datasets. Furthermore, to improve the accuracy of proposed work multiple classifiers such as DT, SVM, and KNN are hybridized. The authors in [8] performs the comparative study of bio-inspired algorithms including SSA, GWO, WO, HHO, used for the FS task of IoT based IDS. The binary and multiclass classification of network traffic is done using the naïve bayes and KNN classifier and their performance are compared against the BoT-IoT dataset.

3. METHODS AND MATERIALS

3.1. Improved Salp Swarm Algorithm (ISSA)

The ISSA is an upgraded form of the native salp swarm algorithm (SSA) proposed by [9] [10]. The SSA imitate the bevy nature of salps in the sea, where they form a chain-

like structure. The salps are grouped as leader and followers. The front salp in the chain serves as the leader, while the other salps in the chain are the followers. Leader location S_n^1 is estimated using a specific equation, which is as follows:

$$S_n^1 = FP_j - rn_1[(UpB - LwB)rn_2 + LwB] \;\; if\; rn_3 < 0.5 \quad (1)$$

$$S_n^1 = FP_j + rn_1[(UpB - LwB)rn_2 + LwB] \;\; if\; rn_3 \geq 0.5 \quad (2)$$

where, FP_j is the food position. LwB is lower bound. rn_2, and rn_3 are the random numbers. UpB is upper bound. rn_1 is calculated as $rn_1 = 2e^{-\left(\frac{4it}{Max_{it}}\right)}$ where, it is current iteration. Max_{it} is maximum iteration. Moreover, the follower location is estimated using the below equation-

$$S_n^m = \frac{1}{2}(S_n^m - S_n^{m-1}) \quad (3)$$

where, S_n^m is the m[th] follower in n[th] dimension.

In ISSA, the opposition based learning is employed in the population initialization phase of SSA for improving the population diversity, and local search boost the exploitation phase of SSA for determining the best solution.

3.2. Harris Hawk Optimization (HHO)

This algorithm introduced in reference [11], emulates the hunting behavior of hawks. In order to capture prey, hawks employ various chasing strategies depending on the scenario and the prey's evasive behaviour. The HHO execution includes two phases: exploration, and exploitation. During exploration, the hawks perform a global search for prey, searching for specific locations and constantly observing their environment. The update of the hawk's position is dependent on either their family members or random searches in the population area. This process can be mathematically modelled as follows:

$$H(it+1) = \begin{cases} H_{rnd}(it) - rn_3|H_{rnd}(it) - 2rn_4\, H(it)| \;\; if\; q \geq 0.5 \\ H_{pr}(it) - H_{mean}(it) - rn_5\big(LwB + rn_6(UpB - LwB)\big)\; if\; q < 0.5 \end{cases} \quad (4)$$

where $H(it)$ is the current hawk position, $H(it+1)$ is the hawk position in next iteration. $rn_3, rn_4, rn_5, rn_6, and\; q$ are all random numbers. The target position that the hawks are trying to reach is denoted as $H_{pr}(it)$. $H_{mean}(it)$ is the hawks mean position, which is calculated as

$$H_{mean}(it) = \sum_{i=1}^{N} \frac{H_i(it)}{N} \quad (5)$$

where N is the total hawk population. The exploration to exploitation transition in HHO is usually described by the following equation-

$$Es_n = 2E_i - \left(1 - \frac{it}{Max_{it}}\right) \quad (6)$$

where Es_n is escaping energy, E_i is initial energy, 'it' is current iteration. Maximum iteration is denoted by Max_{it}. If $Es_n > 1$ then exploration is done otherwise exploitation will be done. During the exploitation, the hawks exploit the prey using different attacking strategies, which includes-

Hard besiege. In this strategy, prey is exhausted and does not escape successfully. Hawk location is updated using the equation below-

$$H(it+1) = H_{pr}(it) - Es_n|\Delta H(it)| \quad (7)$$

$$\Delta H(it) = H_{pr}(it) - H(it) \quad (8)$$

where, $\Delta H(it)$ is the difference between prey position and hawks current position. The target position that the hawks are trying to reach is denoted as $H_{pr}(it)$. $H(it)$ is the current hawk position. Es_n is escaping energy. $H_{pr}(it)$ is prey position.

Soft besiege. In this strategy, prey has enough escaping energy. However, hawks encircle the prey and make them unable to escape. Hawk location is updated using the equation below:

$$H(it+1) = \Delta H(it) - Es_n \left| J H_{pr}(it) - H(it) \right| \quad (9)$$

where $J = 2(1 - rn_7), rn_7$ is the random number. $\Delta H(it)$ is the difference between prey position and hawks current position. J is prey jump strength.

Hard besiege with Progressive Dive. Here, hawks attempt to minimize the distance between their own location and prey location. Hawks location is estimated as

$$H(it+1) = \begin{cases} Y' \; if \; Fit(Y') < Fit(H(it)) \\ Z' \; if \; Fit(Z') < Fit(H(it)) \end{cases} \quad (10)$$

where $Y' = H_{pr}(it) - Es_n \left| J H_{pr}(it) - H_{mean}(it) \right|$ and $Z' = Y' + S \times LFF(Dim)$

where LeF is levy flight function that controls the step size taken by hawks. It is calculated via following equation-

$$LeF = \frac{ur \times \sigma}{|vr|^{\frac{1}{\beta}}}, \sigma = \left(\frac{\Gamma(1+\beta) \times \sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right) \times \beta \times \sin\left(\frac{\beta-1}{2}\right)} \right)^{\frac{1}{\beta}} \quad (11)$$

The value of $\beta = 1.5$ and ur and vr are the random values. Dim is problem dimension. $H_{mean}(it)$ is the hawks mean position calculated using eq. 5. S denotes random vector of size $1 \times Dim$.

Soft besiege with Progressive Dive. Prey has enough escaping energy. Hawks need to find the optimal location to catch the prey. Hawks location is estimated as

$$H(it+1) = \begin{cases} Y \; if \; Fit(Y) < Fit(H(it)) \\ Z \; if \; Fit(Z) < Fit(H(it)) \end{cases} \quad (12)$$

where $Y = H_{pr}(it) - Es_n \left| J H_{pr}(it) - H(it) \right|$ and $Z = Y + S \times LFF(Dim)$

Note: All random numbers used in algorithm implementation lies between 0 and 1.

4. PROPOSED SYSTEM

The IDPS model is shown in Figure 1, is divided into following phases, which include data pre-processing, feature selection, and intrusive traffic detection and classification and prevention. This section describes these phases briefly.

4.1. Data Pre-processing

The different datatypes attribute values prevent the learning algorithm to perform well. Therefore, a transformation of attribute values is done. In this paper, the label encoder function is used to perform this task. Afterward, the normalization of attribute values using standard scaler function is done to make them lie within the range of 0 and 1. Afterward, the problem of data imbalance is resolved using the sampling technique.

The data imbalance issue makes the learning algorithm bend towards the majority class samples.

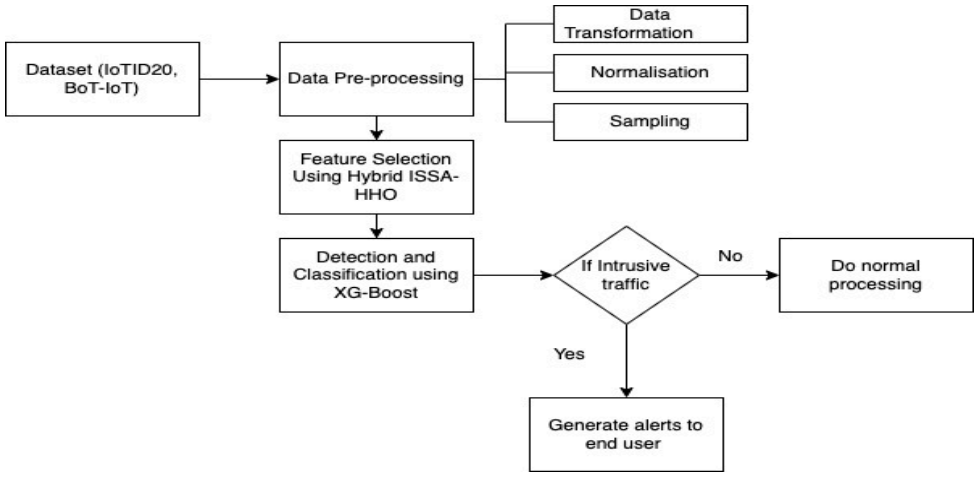

Figure 1. Proposed System Design

4.2. Feature Selection

The feature selection of the pre-processed data is performed using hybridized ISSA-HHO algorithm. The hybridized ISSA-HHO is described in this sub-section. The bio-inspired algorithms are executed in the exploration and exploitation phase. The exploration part of ISSA is modified using HHO algorithm. The trigonometric functions with HHO exploration equations are used to enhance the exploration phase of the ISSA. Afterward, levy flight is used to reform the leader position of the salp. This function controls the step taken by salps. The flow graph of ISSA-HHO feature selection is shown in Figure 2. During exploration, the salps performs the global search without falling into local optima. Here, HHO exploration equations are used to perform the global search. The opposition-based learning [10] is used for the hawks population initialization. Afterward, hawks position is updated as-

$$SH(it+1) = \begin{cases} H_{rnd}(it) - rn_3 \times \sin(rn_0) \times |H_{rnd}(it) - 2rn_4\, H(it)| \;\; if\; q \geq 0.5 \quad 13(a) \\ H_{pr}(it) - H_{mean}(it) - rn_5 \times \cos(rn_0) \times \left(LwB + rn_6(UpB - LwB)\right) if\; q < 0.5 \;\; 13(b) \end{cases} \quad (13)$$

where $rn_0 = \propto -it \times \frac{\alpha}{Max_{it}}$, $\propto = 2$

where $H_{rnd}(it)$ is the random hawks position obtained using opposition based learning. q, rn_3, rn_4, rn_5, and rn_6 are the random numbers. $H_{pr}(it)$ is the prey location. $H(it)$ is current hawks position. $H_{mean}(it)$ is the mean position obtained from equation 5. LwB is lower bound and UpB is the search space upper bound. During exploitation, local search to obtain optimal solution is performed. The LeF is used to control the step size of follower salps. Therefore, the salp follower position is updated as

$$SH_n^m = \frac{1}{2}(SH_n^m - SH_n^{m-1}) + LeF(Dim) \quad (14)$$

where SH_n^m is the current salps position and SH_n^{m-1} is the previous salp positon. $LeF(Dim)$ is levy flight function obtained from equation 11. After that, local search technique [10] is applied to avoid local optima and improve solution.

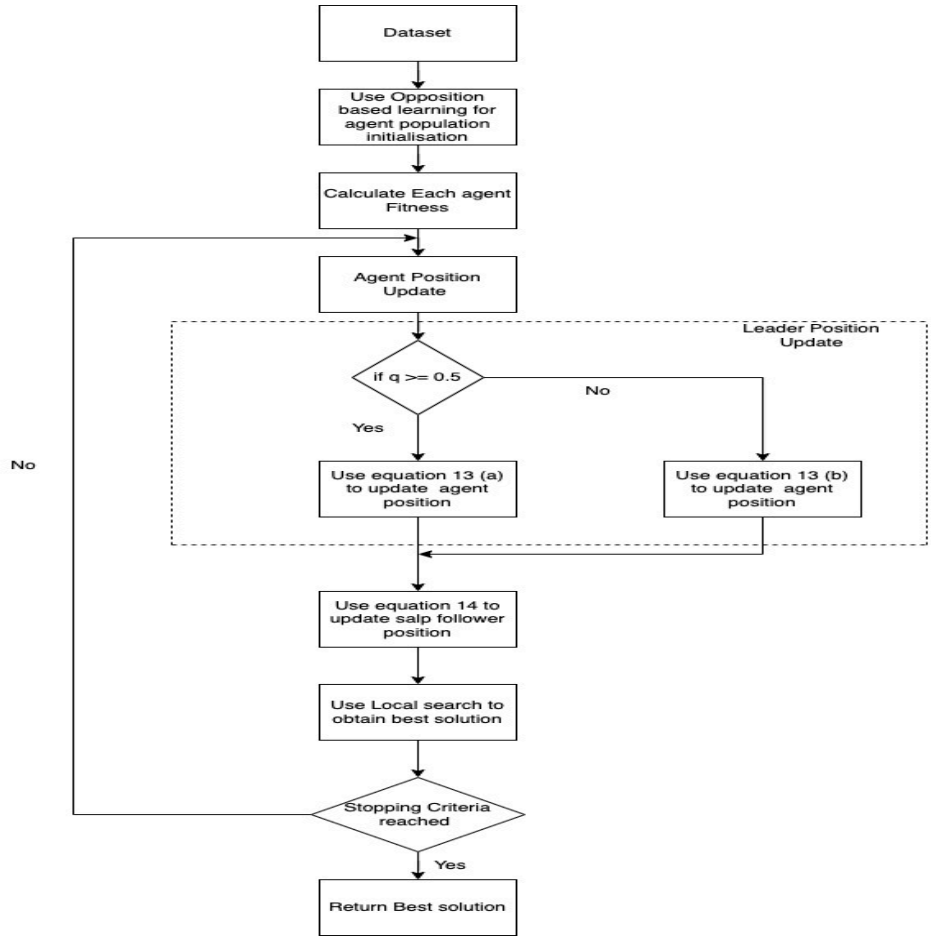

Figure 2. Feature Selection ISSA-HHO

4.3. Classification and Detection

The features selected as provided as input to this phase for classifying intrusive traffic using the XGBoost classifier. This classifier is based on the concept of decision trees and is known for its accuracy, scalability, and efficiency.

4.4. Prevention

In the prevention phase of IDPS, the system first detects and classifies any intrusive or abnormal traffic on the network. If the classified traffic is determined to be normal, it is allowed to proceed through the network without any intervention. However, if the

traffic is classified as intrusive or abnormal, an alarm is generated to notify the end user or security personnel. The user is then prompted to take corresponding actions to prevent any potential damage. This proposed IDPS is placed before the firewall of the end user system. In case of any anomaly detected, rules at the firewall are updated which helps the end user to block the incoming intrusive traffic.

5. IMPLEMENTATION RESULT

The overview of datasets used for the proposed work verification, and implementation result is presented in this section. The implementation of proposed work is done on a MacOS with 8GB memory and 1.6 GHz Dual-Core processor. The python programing language is used for implementation of IDPS.

5.1. Dataset Description

The evaluation of proposed work is done against two IoT based datasets including, BoT-IoT and IoTID20. This section describes these datasets briefly.

BoT-IoT dataset: This dataset is introduced by [12] in 2019. This publically available dataset is created in a realistic IoT environment. The synthetic testbed configuration includes IoT services, analytical tools, and network platform with attacking and normal virtual machines. This labelled dataset with over 72 million records includes 46 features and four attack categories. The table 1 shows the number of BoT-IoT instances (5% of the entire dataset) used for performance evaluation.

Table 1. BoT-IoT Instances

Category	***Instances***	***Category***	***Instances***
DDoS	*1926624*	*Normal*	*477*
DoS	*1650260*	*Theft*	*79*
Reconnaissance	*91082*		

IoTID20 Dataset: This dataset is introduced by authors in [13]. The testbed configuration of this dataset forms a smart home environment which includes two begin devices and all other are attacking devices, connected together via Wi-Fi router. The dataset has over 625,783 data records, includes four attack categories and has 83 features. The table 2 shows the IoTID20 traffic instances used for performance evaluation.

Table 2. IoTID20 Instances

Category	***Instances***	***Category***	***Instances***
Scan	*75265*	*DoS*	*59391*
Mirai	*415677*	*Normal*	*40073*
MITM ARP Spoofing		*35377*	

5.2. Result and Discussion

The proposed IDPS effectiveness is determined by comparing it with other bio-inspired algorithms used as FS method. These algorithms include HHO [11], SSA [9], ISSA [10], particle swarm optimisation(PSO) [14], and bat algorithm (BA) [15].

BoT-IoT dataset- The hybrid ISSA-HHO attains the highest accuracy of 99.8% and BA attains the lowest accuracy of 99.3 % as depicted in Figure 3. Furthermore, least number of features i.e. 8 are selected using ISSA-HHO while the PSO selects the highest number of features as depicted in Figure 4. Computation time taken by proposed ISSA-HHO is lowest in contrast to other algorithms as shown in Figure 5. Although recall for each algorithm is same for BoT-IoT dataset, as shown in Figure 6. The ISSA-HHO converges better than HHO and ISSA as shown in Figure 11.

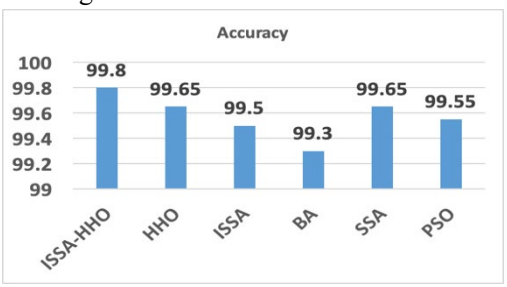

Figure 3. BoT-IoT Accuracy

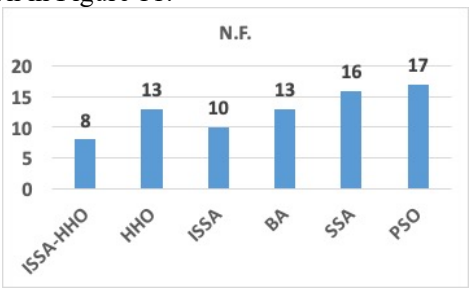

Figure 4. BoT-IoT Number of features

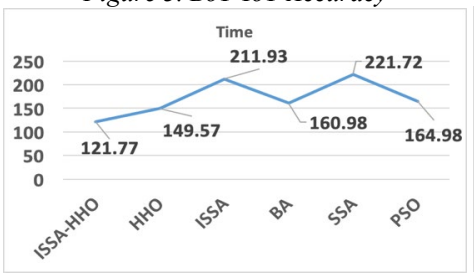

Figure 5. BoT-IoT Time

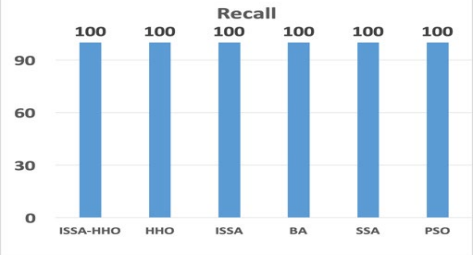

Figure 6. BoT-IoT Recall

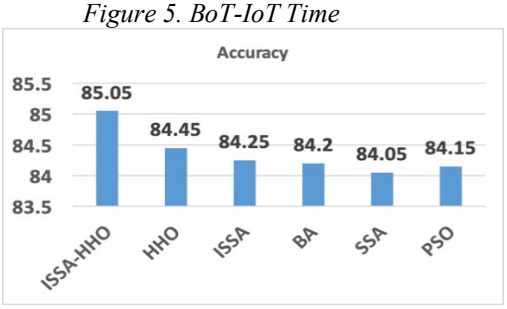

Figure 7. IoTID20 Accuracy

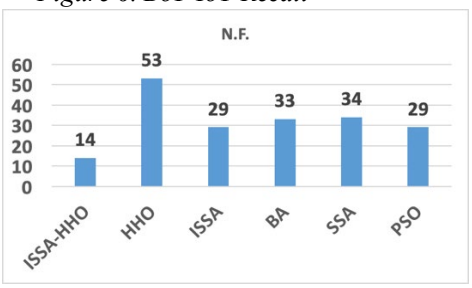

Figure 8. IoTID20 Number of features

IoTID20- From Figure 7 we observed that ISSA-HHO achieves the highest accuracy of 85.05% while SSA attains the lowest accuracy of 84.05%. Moreover, the ISSA-HHO selects the lowest number of features i.e. 14 and the highest number of features i.e. 53 are selected using HHO as depicted in Figure 8. Figure 9 shows that the computation time taken by hybrid ISSA-HHO is comparatively lower than the other algorithms used for FS task. Figure 10 revealed that the recall attained by hybrid ISSA-HHO is higher than ISSA and PSO. Furthermore, the hybrid ISSA-HHO has a better convergence rate than ISSA and HHO as shown in Figure 12.

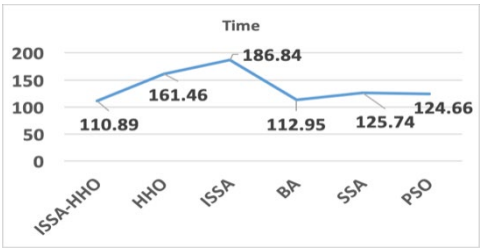

Figure 9. IoTID20 Time

Figure 10. IoTID20 Recall

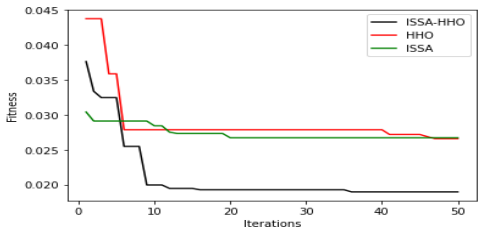

Figure 11. Convergence Curve BoT-IoT

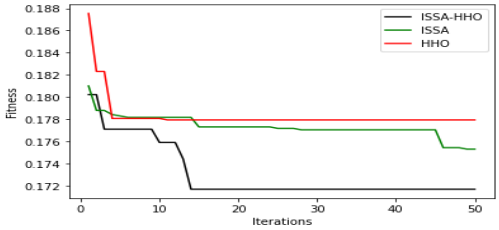

Figure 12. Convergence Curve IoTID20

6. CONCLUSION

The paper introduces a new approach for FS in an IoT-based IDPS by combining the improved salp swarm algorithm (ISSA) and the harris hawk optimization (HHO) algorithm. The proposed technique is then used to identify intrusive traffic using the XGBoost classifier. To evaluate the proposed system, two IoT-based datasets, BoT-IoT and IoTID20, are used. The hybridized FS method is compared against other bio-inspired algorithms such as HHO, ISSA, BA, SSA, and PSO. The outcomes indicate that the proposed system achieves superior detection accuracy with minimal computation time and selected features. Moreover, the hybridized ISSA-HHO-XGBoost exhibits a more rapid convergence rate than the original ISSA and HHO. In essence, the paper aims to tackle the mounting security concerns associated with IoT devices by presenting an efficient and effective FS method for IoT-based IDPS.

REFERENCES

[1] Adamopoulos,I., Ilias A., Makris, C.,Stamatiou, Y.C. Intelligent Surveillance Systems on the Internet of Things based on secure applications with the IBM cloud platform. *International Journal on Information Technologies & Security*, vol. 15, no. 2, 2023.

[2] Kumar, R.A., Franklin, J.V., Koppula, N. A Comprehensive survey on metaheuristic algorithm for feature selection techniques. *Materials Today: Proceedings*, 2022.

[3] Kareem, S. S., Mostafa, R. R., Hashim, F. A., El-Bakry, H. M. An effective feature selection model using hybrid metaheuristic algorithms for IoT intrusion detection. *Sensors*, vol. 22, no. 4, 2022, p.p.1396.

[4] Mafarja, M., Heidari, A. A., Habib, M., Faris, H., Thaher, T. Augmented whale feature selection for IoT attacks: Structure, analysis and applications. *Future Generation Computer Systems*, vol. 112, 2020, pp. 18-40.

[5] Louk, M. H., Tama, B. A. PSO-driven feature selection and hybrid ensemble for network anomaly detection. *Big Data and Cognitive Computing*, vol. 6, no. 4, 2022.

[6] Dahou, A., Elaziz, M. A., Chelloug, S. A., Awadallah, M. A., Al-Betar, M. A., Al-qaness, M. A., Forestiero, A. Intrusion detection system for IoT based on deep learning and modified reptile search algorithm. *Computational Intelligence and Neuroscience*, 2022, pp. 1-15.

[7] Naseri, T. S., Gharehchopogh, F. S. A feature selection based on the farmland fertility algorithm for improved intrusion detection systems. *Journal of Network and Systems Management*, vol. 30, no. 3, 2022.

[8] Singh, R., Ujjwal. R. L. Feature selection methods for IoT intrusion detection system: Comparative study. In *Computational Intelligence. Lecture Notes in Electrical Engineering*, 2023.

[9] Mirjalili, S., Gandomi, A. H., Mirjalili, S. Z., Saremi, S., Faris, H., Mirjalili, S. M. Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems. *Advances in Engineering Software*, vol. 114, 2017, pp. 163-191.

[10] Tubishat , M., Idris, N., Shuib, L., Abushariah, M. A., Mirjalili, S. Improved Salp Swarm Algorithm based on opposition based learning and novel local search algorithm for feature selection. *Expert Systems with Applications*, vol. 145, 2020

[11] Heidari, A. A., Mirjalili, S., Faris, H., Aljarah, I., Mafarja, M., Chen, H. Harris hawks optimization: Algorithm and applications. *Future Generation Computer Systems*, vol. 97, 2019, pp. 849-872.

[12] Koroniotis, N., Moustafa, N., Sitnikova, E., Turnbull B. Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset. *Future Generation Computer Systems*, vol. 100, 2019, pp. 779-796.

[13] Ullah, I. Mahmoud, Q. H. A scheme for generating a dataset for anomalous activity detection in IoT networks. In *Canadian Conference on Artificial Intelligence*, 2020.

[14] Kennedy, J., Eberhart, R. Particle swarm optimization. in *Proceedings of ICNN'95 - International Conference on Neural Networks,* Perth, WA, Australia, 1995.

[15] Yang, X., Gandomi, A. H. Bat algorithm: a novel approach for global engineering optimization. *Engineering Computations*, vol. 29, no. 5, 2012, pp. 464-483.

Information about the authors:

Richa Singh is a research scholar in Computer Science and Engineering Department at USIC&T, G.G.S.I.P. University Delhi, India. Her area of interest includes Artificial Intelligence, Network Security, and Internet of Things.

Dr. R.L. Ujjwal is working as an Associate Professor at USIC&T, G.G.S.I.P. University Delhi, India. His area of interest includes Internet of Things, Cloud Computing, and Wireless Mobile Networks.

Remark:

Manuscript received on 07 July 2023 and is accepted after double-blind reviewing to take part in the 37th International Conference on Information Technologies (InfoTech-2023), IEEE conference, Rec. # 58664, Section F: "Technological Aspects of e-Governance and Privacy" and has not been published in full text elsewhere.