International Journal of Research and Publication
Issue 5:Vol. 2 (May, 2023)

Enhancing Intrusion Detection Systems for Supply Chain Attacks Using Optimized Machine Learning Models: A Case for BPSO-XGBoost

Solomon Goldman Olumba[1], Frederick Oscar[2], Ugboaja Samuel Gregory[3], Charles C. Anyim[4], Luka Ngoyi[5], Ifeyinwa Nkemdilim Obiokafor[6], Peretula Akhamie[7] Anthony Olusanya Fakoya[8], Adeleye Olufemi[9], Philip Ugbede Ojo Onuche[10]

[1]School of Built Environment, Engineering and Computing, Leeds Beckett University
[1]S.olumba3479@student.leedsbeckett.ac.uk
[2]Adioo technology, Fredeoscar90@gmail.com
[3]Michael Okpara University of Agriculture Umudike ugboaja.samuel@mouau.edu.ng
[4]Geography and GIS, University of South Carolina, Columbia, USA
[4]ugochukwu.udonna.okonkwo@gmail.com
[5]University of Zambia. School of Engineering luka.ngoyi@unza.zm
[6]Computer Science Technology, Anambra State Polytechnic, Mgbakwu
[6]ifykems@gmail.com
[7]Management Information Systems, Southern Illinois University Edwardsville
[7]tutuakhamie@gmail.com
[8]ICT/ Computer Science Departments, Metallurgical Training Institute (MTI) Onitsha.
[8]sanyafakoya2@gmail.com
[9] School of mathematics and computer science, University of Wolverhampton
[9]samjwbox@gmail.com
[10]Department of Business Administration, Founders Hall, Southern Illinois University Edwardsville
[10]onucheugbedeojo240@gmail.com

Abstract—Cybersecurity threats, including the increase of sophisticated supply chain attacks, continue to escalate in complexity and volume, necessitating the development of robust and efficient intrusion detection systems (IDS). This study presents an enhanced intrusion detection model that combines Binary Particle Swarm Optimization (BPSO) for feature selection with XGBoost for classification. By leveraging BPSO's optimization capabilities and XGBoost's ensemble learning strengths, the proposed BPSO-XGBoost model demonstrates superior performance in identifying malicious activities in network traffic. The methodology is validated using real-world cybersecurity datasets, achieving significant improvements in metrics such as sensitivity, F1 score, and AUC-ROC compared to traditional machine learning models. These results emphasize the model's potential to strengthen IDS frameworks and mitigate advanced cyber threats effectively.

Index Terms—Cybersecurity, Intrusion Detection Systems, Machine Learning, Feature Selection, BPSO-XGBoost.

I Introduction

Intrusion Detection Systems (IDS) are vital components of modern cybersecurity frameworks, designed to identify and mitigate threats within network environments [18]. The dynamic and evolving nature of cyber-attacks, such as distributed denial-of-service (DDoS) attacks, malware infiltration, and phishing and supply chain attacks, necessitates advanced machine learning techniques capable of distinguishing between benign and malicious network behaviors with high accuracy.

Building upon the foundation of the BPSO-XGBoost model as demonstrated in [1], this study aims to adapt and enhance its capabilities specifically for IDS applications, including the detection of supply chain attack patterns. BPSO optimizes feature selection, reducing data dimensionality and improving computational efficiency, while XGBoost leverages ensemble learning to maximize classification performance. Together, these techniques create a robust model suited to address the challenges of modern network security, with a direct focus on preventing vulnerabilities introduced through supply chain attacks.

II Related Works

In this section, we present the numerous studies on the application of machine learning models in the malware attack detection process. In recent times, the integration of deep learning techniques into cybersecurity frameworks has gained significant attention, particularly in sectors vulnerable to cyber threats, such as shipping. Existing frameworks have primarily focused on traditional security measures, often lacking the proactive capabilities necessary to anticipate and mitigate potential attacks. For instance, studies have demonstrated the effectiveness of deep learning in anomaly detection and threat prediction [1–5], which are critical for safeguarding maritime operations. However, many of these approaches do not fully address the unique challenges faced by shipping enterprises, such as the dynamic nature of maritime environments and the increasing sophistication of cyber threats. The authors in [6], builds upon the foundational work in this area by proposing an attack-aware cybersecurity framework that leverages deep learning to enhance the resilience of shipping enterprises against evolving cyber threats, thereby filling a crucial gap in the current literature. The study in [7], investigated the effectiveness of generalized entropy (GE) and the generalized information divergence (GID) based on information theory. Results from their studies show that the GID and GD are very effective in detecting DDoS attacks. The work in [8], presents compelling evidence on robust attack detection approaches by

investigating the use of deep learning approaches for cyber-physical systems in supply Chain 4.0 security. Although the study indicated that ML is efficient in detecting attacks. The authors primarily evaluated the proposed approach using real-world semiconductor production factory data, thus overlooking wider applications.

To contextualize the study or different cyber-attacks, the study in [9] investigated various types of cyber-attacks targeting concentration sensors in chemical engineering processes. While the study proposed ML-based detection methods, it did not thoroughly explore the scalability or generalizability of the proposed techniques beyond the chemical engineering domain. The paper builds upon the approach concerning cyber-physical systems (CPSs) and Distributed Denial of Service (DDoS)

Feature Group	Count	Description
IMAGE DOS HEADER	19	Metadata about the MS-DOS header, including *e magic* and *e l fanew*.
FILE HEADER	7	File structure details such as Machine, NumberOfSections, and Characteristics.
OPTIONAL HEADER	29	Execution and memory layout details like ImageBase, Subsystem, and SizeOfImage.
Derived Features	15	Computed attributes enhancing analysis.
Total Features	**69**	Combination of raw and derived features.
Target Variable	**1**	Binary class label: 0 (benign) or 1 (malware).

attack detection [10, 18]. The PCA-BSO feature selection strategy significantly improves the detection capabilities used in the work. However, the paper could not take into consideration the discussion on computational complexity and scalability of the PCA-BSO, which are essential in real-time application in a large-scale CPS environment. The study further evaluated the performances of different ML and deep learning approaches. Building on previous works by [11] on LSTM-based approaches, focus was made on multivariate time series. The study in [11], further highlighted the efficiency of their proposed method in ECG anomaly detection. However, LSTM models have limitations such as complex hyperparameter tuning and sensitivity to noise [12].

III Methodology

A. Dataset

The dataset used in this study is derived from the ClaMP dataset, originally designed for malware classification using header field values of Portable Executable (PE) files. It consists of 5,184 labelled samples, divided into benign and malicious categories, with 69 features comprising 54 raw attributes and 15 derived features. Key feature groups include IMAGE _DOS HEADER, FILE _HEADER, and OPTIONAL HEADER, providing detailed metadata for analysis. This dataset was selected for its comprehensive representation of executable file behaviors, enabling robust modelling of malware detection systems applicable in supply chain cybersecurity. The features of the ClaMP dataset is shown in Table I.

Table: I Key Features of the ClaMP Dataset

B. Preprocessing

• **Data Cleaning and Transformation:** Initial preprocessing involves handling missing values, normalizing continuous features, and encoding categorical variables.

• **Feature Selection with BPSO:** Binary Particle Swarm Optimization is applied to identify the most relevant features, enhancing model performance and reducing computational requirements.

C. Model Training and Evaluation

The data for the experiment is split into 60/20/20, training, validation and testing ratio. The optimized feature set is input into the XGBoost classifier. Hyperparameters are fine-tuned using grid search and cross-validation to achieve optimal performance. The models are evaluated using metrics such as sensitivity, F1 score, and AUC-ROC, alongside computational efficiency metrics.

D. Binary Particle Swarm Optimization (BPSO) as an Optimization Technique

Binary Particle Swarm Optimization (BPSO) is an adaptation of the traditional Particle Swarm Optimization (PSO) algorithm, tailored for discrete binary spaces. It is widely employed for feature selection [13, 14], optimizing the subset of features to improve model performance while reducing computational complexity.

E. Mathematical Formulation of BPSO

BPSO models a set of candidate solutions as particles in a binary search space. Each particle i is characterized by:

- **Position:** $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{id}]$, where $x_{ij} \in \{0,1\}$ represents the binary decision to include (1) or exclude (0) feature j.
- **Velocity:** $\mathbf{v}_i = [v_{i1}, v_{i2}, \ldots, v_{id}]$, which influences the probability of a bit flipping between 0 and 1.

As shown in Equation (1.) velocity of each particle is updated using:

$$v_{ij}(t+1) = \omega v_{ij}(t) + c_1 r_1 [p_{ij}(t) - x_{ij}(t)] + c_2 r_2 [g_j(t) - x_{ij}(t)] \text{ ---------- (1),}$$

where:

- ω is the inertia weight, balancing exploration and exploitation.
- c_1 and c_2 are cognitive and social coefficients, respectively.
- r_1, r_2 are random values in [0,1].
- p_{ij} is the particle's best-known position.
- g_j is the global best position among all particles.

The Equation (2.) and(3) demonstrate that the binary position update is determined using a sigmoid function:

$$s\left(v_{ij}(t+1)\right) = \frac{1}{1+e^{-v_{ij}(t+1)}} \text{ ---------- (2)}$$

followed by:

$$x_{ij}(t+1) = \begin{cases} 1 & if\ r \le s(v_{ij}(t+1)), \\ 0 & otherwise, \end{cases} \text{ ---------- (3)}$$

where r is a random value in [0,1].

BPSO selects the optimal feature subset by maximizing a fitness function, typically based on classification accuracy or a combination of accuracy and feature reduction.

F. XGBoost as a Classification Technique

XGBoost (Extreme Gradient Boosting) is a high performance, scalable machine learning algorithm based on gradient-boosted decision trees [15, 16]. It excels in structured data tasks, offering robust classification capabilities.

G. Mathematical Formulation of XGBoost

XGBoost builds an additive model by iteratively improving a base model. The predicted output at iteration t is shown in Equation (4.) below:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t+1)} + f_t(x_i) \text{ ---------- (4)}$$

where f_t is a decision tree added at iteration t, and is the prediction for sample i.

The objective function for minimization is indicated in Equation (5):

$$\mathcal{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i^{(t)}) \text{ ---------- (5)}$$

where:

• ℓ is a differentiable loss function (e.g., log-loss for classification).

• $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ is a regularization term to prevent overfitting.

For a single tree, the optimal weight w_j for leaf j is calculated in Equation (6) as:

$$w_j = -\frac{\sum_{i\in I_j} g_j}{\sum_{i\in I_j} h_i + \lambda}, \text{ ---------- (6)}$$

where g_i and h_i are the first and second order gradients of the loss function.

H. The Proposed BPSO-XGBoost Model

In our research, we have strategically built upon the foundational framework presented in [1], which integrates Binary Particle Swarm Optimization (BPSO) with the advanced classification capabilities of XGBoost. This combination effectively addresses critical challenges in intrusion detection systems (IDS) by optimizing feature selection and enhancing classification accuracy. By leveraging BPSO's feature selection strengths and XGBoost's powerful classification algorithms, this methodology has set a new standard for improving IDS model efficiency and precision, and we validate upon this framework to further refine and extend its capabilities.

Our proposed framework refines and extends this integration to deliver a more robust and scalable solution. While the work in [1] provided a solid foundation, our innovation further develops these techniques to address the increasing complexity of cyber threats. By enhancing the synergy between BPSO and XGBoost, we have developed a model that not only improves detection capabilities but also strengthens the real time mitigation of advanced cyber-attacks.

This work is particularly relevant in addressing the growing sophistication of cyber threats, which demand both high precision and scalability from IDS solutions. Our framework overcomes the limitations of traditional intrusion detection methods, offering a superior alternative for safeguarding critical infrastructure and sensitive systems from cyber adversaries.

Our contributions represent a significant advancement in cybersecurity, showcasing the transformative potential of combining optimization techniques with machine learning to enhance the effectiveness of security systems. The Algorithm 1 depicts a pictorial representation of the propose BPSO-XGBoost integration process.

I. Integration Process

1. **Feature Selection:** BPSO optimizes the feature set by selecting a subset that maximizes the fitness function, balancing classification accuracy and feature reduction.
2. **Classification:** The selected feature subset is entered into the XGBoost classifier. Hyperparameters are finetuned using grid search and cross-validation for optimal performance.
3. **Combined Objective:** The overall objective is to maximize the model's performance metrics (e.g., sensitivity, F1 score, AUC-ROC) while minimizing computational overhead.

Mathematically, let F denote the full feature set and S denote the subset selected by BPSO.

The objective of the BPSO-XGBoost model is to optimize:

$\mathscr{F}$ **(S)** = Maximize Accuracy (XGBoost(S))−α‖S‖,

where α is a regularization parameter that penalizes larger subsets.

IV Pseudocode for BPSO-XGBoost

Algorithm 1 BPSO-XGBoost Framework

1: **Input:** Feature set **F,** Population size N, Max iterations T_{max}, Inertia weight ω, Cognitive coefficient c_1, Social coefficient c_2, XGBoost hyperparameters H.
2: **Output:** Optimized feature subset S^*, Trained XGBoost model.
3: Initialize particle positions x_i and velocities v_i for i = 1,2,...,N.
4: Evaluate fitness f_i = Accuracy (XGBoost(S_i))−$\alpha\|S_i\|$.
5: Set personal best $\mathbf{p}_i = \mathbf{x}_i$ and global best **g.**
6**: for** t = 1 to T_{max} **do**
7: **for** each particle i **do**
8: Update velocity: $v_{ij}(t + 1) = \omega v_{ij}(t) + c_1 r_1[p_{ij}(t)-x_{ij}(t)]+c_2 r_2[g_j(t)-x_{ij}(t)]$.
9: Update position: $x_{ij}(t+1)$ = sigmoid($v_{ij}(t+1)$).
10: **end for**
11: Evaluate fitness for each particle and update $\mathbf{p}_i$ and **g**.
12: **if** Stopping criterion met **then**
13: **Break.**
14: **end if**
15: **end for**
16: **Feature Selection:** Use **g** to select $\mathbf{S}^*$.
17: **Model Training:** Train XGBoost on $\mathbf{S}^*$ with **H**.
18: **Return:** Optimized feature subset $\mathbf{S}^*$ and trained model.

V Results and Discussion

In this section, we analyse the performances of the proposed BPSO-XGBoost over the baseline models.

A. Sensitivity Performance Analysis

Fig. 1 shows the sensitivity performances of the proposed BPSO-XGBoost over the state-of-the-art (SOTA)) models.

The BPSO-XGBoost model attains the highest sensitivity of **0.95,** outperforming all baseline models. The proposed model exhibits its superior ability to detect true positives effectively, which is very important in intrusion detection systems. The Random Forest and XGBoost models obtain sensitivities of 0.88 and 0.85, respectively, showcasing their reliability lag the proposed model's performance.

B. F1 Score Performance Analysis

Analysis: demonstrated in Fig 2., the BPSO-XGBoost model achieve the highest F1 score of **0.86,** demonstrating its ability to establish a balance between precision and recall. This out-performance highlights

F1-scores for Different Models at 200K Samples of DDoS

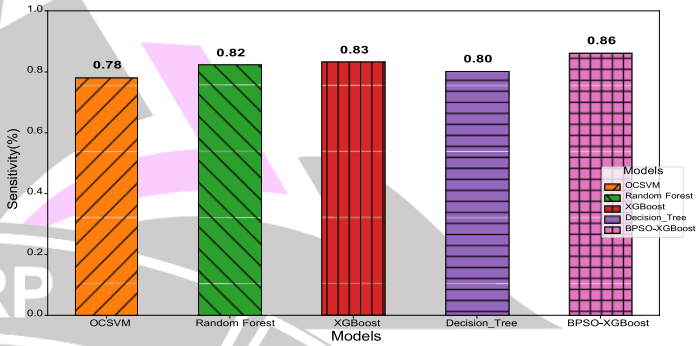

Fig. 1: Hyperparameter Optimization Comparison

F1-scores for Different Models at 200K Samples of DDoS

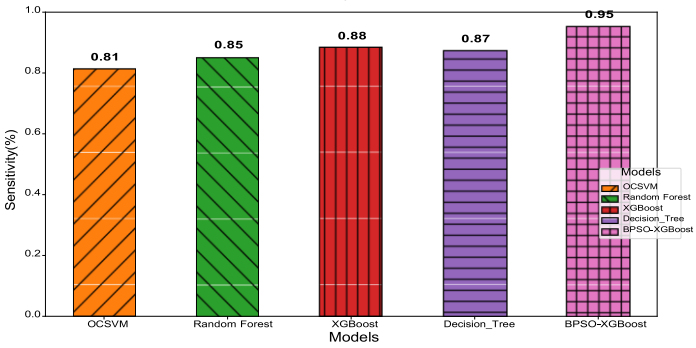

Fig. 2: Hyperparameter Optimization Comparison

the efficiency of BPSO in boosting feature selection and improving the classification accuracy of the XGBoost model.

C. AUC-ROC Performance Analysis

In this scenario, the AUC-ROC results, which demonstrates the model's ability to distinguish between classes, are presented in Fig. 3.

The AUC-ROC scores shows that the BPSO-XGBoost model obtains the highest value of **0.97,** indicating its superior classification ability. Though The XGBoost model exhibits a good performs with an AUC-ROC of 0.96, while Decision Tree and Random Forest models score 0.93, the all fall behind the proposed model in performance. OCSVM attains 0.95 auc performance respectively, indicating its comparatively lower capacity for distinguishing between malicious and benign network traffic.

D. Combined Performance Insights

The combined analysis of sensitivity, F1 scores, and AUC-ROC metrics shows the robustness and reliability of the BPSO-XGBoost model. Its ability to performance across all evaluation metrics underscores its suitability for real-world intrusion detection systems. The integration of BPSO into the XGBoost framework enhances feature selection, reduces noise, and improves classification performance, making it a powerful tool for modern cybersecurity applications.

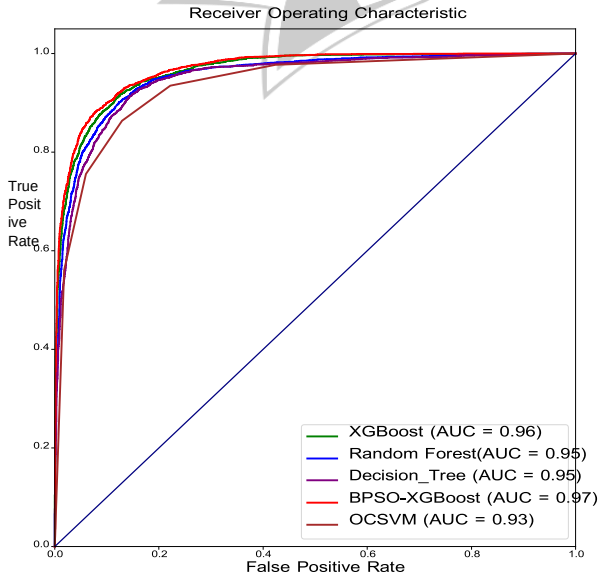

Fig. 3: Hyperparameter Optimization Comparison

VI Conclusion

The proposed BPSO-XGBoost model significantly advances the capabilities of intrusion detection systems by combining optimized feature selection with high-performance classification. The model consistently outperforms baseline methods in sensitivity, F1 score, and AUC-ROC metrics, demonstrating its effectiveness in accurately identifying malicious activities. These results validate the model's potential to enhance IDS frameworks, providing a scalable and efficient solution for combating advanced cyber threats. Future work will explore its application to encrypted traffic and deployment in live network environments.

References

[1] E. Eziama, S. Ahmed, S. Ahmed, F. Awin, and K. Tepe, "Detection of adversary nodes in machine-to-machine communication using machine learning based trust model," in *2019 IEEE international symposium on signal processing and information technology (ISSPIT)*, pp. 1–6, IEEE, 2019.

[2] E. Eziama, L. M. Jaimes, A. James, K. S. Nwizege, A. Balador, and K. Tepe, "Machine learning-based recommendation trust model for machine-to-machine communication," in *2018 IEEE International Symposium on Signal*

Processing and Information Technology (ISSPIT), pp. 1–6, IEEE, 2018.

[3] E. Eziama, K. Tepe, A. Balador, K. S. Nwizege, and L. M. Jaimes, "Malicious node detection in vehicular ad-hoc network using machine learning and deep learning," in *2018 IEEE Globecom Workshops (GC Wkshps)*, pp. 1–6, IEEE, 2018.

[4] E. Eziama, F. Awin, S. Ahmed, L. Marina Santos Jaimes, A. Pelumi, and D. Corral-De-Witt, "Detection and identification of malicious cyber-attacks in connected and automated vehicles' real-time sensors," *Applied Sciences*, vol. 10, no. 21, p. 7833, 2020.

[5] E. Eziama, *Emergency Evaluation in Connected and Automated Vehicles*. PhD thesis, University of Windsor (Canada), 2021.

[6] Z. Wu, J. Wang, Q. Shi, J. Zhang, J. Liu, and X. Zhang, "An attack-aware shipping enterprise cybersecurity framework based on deep learning," in *2023 11th International Conference on Information Systems and Computing Technology (ISCTech)*, pp. 115–119, IEEE, 2023.

[7] S. Behal and K. Kumar, "Detection of ddos attacks and flash events using information theory metrics–an empirical investigation," *Computer Communications*, vol. 103, pp. 18–28, 2017.

[8] S. S. Abosuliman, "Deep learning techniques for securing cyber-physical systems in supply chain 4.0," *Computers and Electrical Engineering*, vol. 107, p. 108637, 2023.

[9] S. Parker, Z. Wu, and P. D. Christofides, "Cybersecurity in process control, operations, and supply chain," *Computers & Chemical Engineering*, p. 108169, 2023.

[10] H. D. Nguyen, K. P. Tran, S. Thomassey, and M. Hamad, "Forecasting and anomaly detection approaches using lstm and lstm autoencoder techniques with the applications in supply chain management," *International Journal of Information Management*, vol. 57, p. 102282, 2021.

[11] P. Malhotra, L. Vig, G. Shroff, P. Agarwal, *et al.*, "Long short-term memory networks for anomaly detection in time series.," in *Esann*, vol. 2015, p. 89, 2015.

[12] J. Brownlee, *Long short-term memory networks with python: develop sequence prediction models with deep learning*. Machine Learning Mastery, 2017.

[13] J. Zhu, J. Liu, Y. Chen, X. Xue, and S. Sun, "Binary restructuring particle swarm optimization and its application," *Biomimetics*, vol. 8, no. 2, p. 266, 2023.

[14] G. Deliorman and D. Inan, "Binary particle swarm optimization as a detection tool for influential subsets in linear regression," *Journal of Applied Statistics*, vol. 48, no. 13-15, pp. 2441–2456, 2021.

[15] J. Deng, L. Cheng, H. Yuan, K. Zheng, X. Li, and Q. Li, "An online detection system for ldos attack based on xgboost," in *2023 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pp. 1083– 1088, IEEE, 2023.

[16] S. S. Dhaliwal, A.-A. Nahid, and R. Abbas, "Effective intrusion detection system using xgboost," *Information*, vol. 9, no. 7, p. 149, 2018.

[17] I. N. Obiokafor, "Strategies to Mitigate Cyber Identity Theft in Africa's Digital Transformation." *JASSD-Journal of African Studies and Sustainable Development* 7.4 2024.

[18] I. N. Obiokafor, and F. C. Aguboshim, "Cybersecurity Strategies for Safeguarding Smart Ecosystem Infrastructure: A Narrative Review." *ANSPOLY Journal of Advanced Research in Science & Technology (AJARST)* 1.1 49-64. 2024.