Journal of Information Systems Engineering and Management
2025, 10(11s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Optimized Feature Selection techniques for Distributed Intrusion Detection System (DIDS) in IoT Environment

B.Karthikeyan, (corresponding author), Dr. K. Kamali, Dr. R. Manikandan
Assistant Professor, Department of English, Annamalai University. Trichy, Tamilnadu.karthikeyanphd@yahoo.com
Assistant Professor, Dept of Computer science,Annamalai University. Trichy, Tamilnadu.
Kamaliaucse2006@gmail.com
Associate Professor, Dept of Computer science,Annamalai University. Trichy, Tamilnadu.
rmkmanikandan1111@gmail.com

ARTICLEINFO

Received: 18 Nov 2024
Revised: 24 Dec 2024
Accepted: 15 Jan 2025

ABSTRACT

Introduction: In Internet of Things (IoT), optimal data set management and feature learning are the important problems that affect the attack detection's accuracy.

Objectives: To select the optimal features in the dataset, using advanced optimization techniques and classify the data as normal or anomaly,using ML based algorithms.

Methods: In this paper, we propose to design anoptimized feature selection technique for Distributed Intrusion Detection System (DIDS)in IoT environments. During the preprocessing phase, t-distributed Stochastic Neighbor Embedding (t-SNE) is applied for data exploration and visualizing the high-dimensional data and Principal Component Analysis (PCA) technique is applied for dimensionality reduction.Then, for selecting the optimal features from the preprocessed dataset, the Improved Gravitational Search Algorithm (IGSA) is applied. Finally, for classifying the data as normal or anomaly, the XGBoost classifier is applied.

Results: Experimental results show that the optimized XGBoost classifier attains highest accuracy and F1-score values, when compared to the other classifiers

Conclusion: The proposed DIDS thus protects the IoT networks from external attacks quickly and effectively.

Keywords: Internet of Things (IoT), Distributed Intrusion Detection System (DIDS), Optimized feature selection, Improved Gravitational Search Algorithm (IGSA) ,XGBoost classifier

INTRODUCTION

In today's rapidly developing digital environment, every device is connected with the physical world using the IoT. IoT is progressively considered as the "Internet of Everything" (IoE) because it includes all types of smart devices [1]. According to a global market survey, by 2025, the number of connected IoT devices is estimated to be 21.5 billion. In addition, IoT systems usually store and maintain data in a distributed manner, instead of depending on the very centralized method of collecting storage and processing resources in huge data centers [2].The absence of dedicated anomaly detection systems and robust security measures in heterogeneous IoT networks makes them susceptible to numerous attacks like spoofing, data leakage, and denial of service (DoS/DDoS). These susceptibilities can cause severe consequences such as system disruptions, hardware damage, and even physical harm. [3].

Efficient Intrusion Detection Systems (IDSs) are required for protecting IoTsmart devices while lessening resource consumption. A Distributed IDS (DIDS) for ioT environment is mainly efficient since it can detect abnormal behaviour in a component by using the cooperation between different IoT devices[4]. Owing to the unique characteristics of IoT networks, implementing DIDS in IoT environments presents several challenges [5].

ML and DL approaches have been progressively proposed to recognize and alleviate security threats. However, conventional ML approaches did not have optimal data set management and feature learning, which can weaken the attack detection's accuracy. In high-dimensional data generated by thousands of IoT sensors and devices, managing inappropriate features can cause overfitting, making decisions based on more training time and noise [6]

Conventional anomaly detection systems are ineffectual in IoT ecosystems since the range of normal behaviors shown by devices is much wider and more dynamic when compared to that in conventional IT environments. The attack detection's accuracy is significantly influenced by challenges in data set management and optimal feature learning. But existing IDSs in the literature did not concentrate on this issue.

OBJECTIVES

- To select the optimal features in the dataset, using advanced optimization techniques.
- To classify the data as normal or anomaly,using ML based algorithms.

RELATED WORKS

Bakhsh et al. [7] proposed an adaptive IDS and Prevention System (IDPIoT) to reinforce security when the number of connected devices grows.Qaddoura et al. [8] proposed a novel three-stage approach for IDS, which includes oversampling, clustering and data reduction, and classification using a Single Hidden Layer Feed-Forward Neural Network (SLFN).

Sohail et al. [9] proposed a Multi-tiered ANN Model for IDS (MAMID), which is a scalable solution for optimal hyperparameter selection for achieving high accuracy in identifying security attacks.Wang et al. [10] proposed a Transformer-based IoT Network IDS (NIDS) that learns attack behaviours from different data types generated in heterogeneous IoT environments.

PROPOSED METHODOLOGY

Overview

In this work, we propose to design a DIDS for IoT, using ML models and optimization techniques. The KDD cup dataset will be used in the training process. Initially, during pre-processing phase, for data exploration and visualizing the high-dimensional data, t-SNE and PCA techniques are applied which are unsupervised non-linear dimensionality reduction techniques. For selecting the optimal features from the pre-processed dataset, the IGSA algorithm is applied. For classifying the data as normal or anomaly, the XGBoost classifier is applied.

In this work, the KDD cup dataset [11] is used for training. It contains the following 42 features with 494021 records. The target class "intrusion_type" contains 23 output labels.

Phase 1 Preprocessing

For data exploration and visualizing the high-dimensional data, t-SNE and PCA techniques are applied which are unsupervised non-linear dimensionality reduction techniques.

Phase 2: Optimal Feature Selection

For selecting the optimal features from the pre-processed dataset, the Improved Gravitational Search Algorithm (IGSA) [13] is applied.

The steps involved in this algorithm are as follows:

1. Initialize the adaptive GSA to create the initial particle swarm.

2. Define parameters comprising the maximum number of iterations NC_{Max}, swarm size N, maximum distance Rp_{max}, search space dimension XDim, minimum distance Rp_{min}, gravitational constant, attenuation rate, constant value, and other relevant parameters.

3. Assess the particle boundaries within the population and calculate the fitness values for all particles.

The following equations are used to calculate best(t) and worst(t).

$$\begin{cases} B(t) = \max fit(t), i \in \{1,2, \dots N\} \\ W(t) = \min fit(t), i \in \{1,2, \dots N\} \end{cases} \quad (1)$$

4. Obtain the inertial mass $Z_i(t)$ of the particles on the basis of B(t) and W(t) from

Equation (6).

$$\begin{cases} z_i(t) = \frac{fit_i(t) - w(t)}{b(t) - w(t)} \\ Z_i(t) = \frac{z_i(t)}{\sum_{j=1}^{N} m_j(t)} \end{cases} \quad (2)$$

5: Update the gravitational constant R as per Equation 7.

$$R(t) = \frac{R_0}{1+e^{\beta(t-t_c)/T}} 0 \leq t_c < T \quad (3)$$

6: Compute the distances between particles as per Equations (4)-(5),

$$\text{Euclidean distance } D_{ij}(t) = ||X_i(t) - X_j(t)||_2 \quad (4)$$

$$\text{Population Density PD} = \frac{1}{N}\sum_{i=1}^{N} d_i(t) \quad (5)$$

where N, D, and d_i indicate the number of particles, dimensionality, and average distance between the i^{th} particle and other particles.

Fixed distance is computed as follows:

$$\text{FD (PD)} = \{FD_{min} + (FD_{max} - FD_{min})e^{1-\frac{1}{PD}} \; PD < 1 \quad (6)$$

$$\text{FD (PD)} = \{FD_{min} + (FD_{max} - FD_{min})e^{1-PD} \; PD \geq 1$$

Where PD is population density, FD_{min} *and* FD_{max} are the minimum and maximum values of the fixed distance respectively,

7: : Compute the gravitational and resultant forces (GF) around the particles, which is given by:

$$GFGF_{ij}^{k}(t) = G(t)\frac{IM_{ai}(t) \times IM_{aj}(t)}{F_{ij}^{FD(PD)}(t) + \epsilon}(x_j^k(t) - x_i^k(t)) \quad (7)$$

Where ε = a very small constant, $IM_{aj}(t)$ = inertial mass of the action object j, $IM_{ai}(t)$ = inertial mass of the action object i

8: Calculate the particles' acceleration as per Equation (8).

$$ACC_i(t) = F_i(t)/M_i(t) \quad (8)$$

9: Update the particles' speeds and positions as per Equations (9)-(12).

$$VE_i^k(t+1) = rand_i \times VE_i^k + ACC_i^k)(t) \quad (9)$$

The adaptive position update is given by:

$$x_i^k(t+1) = \mu \times x_i^k(t) + \gamma \times v_i^k(t+1) \quad (10)$$

where $\mu = e^{-dim*(t/T_{\max})w}$ (11)

$$\gamma = 1 - \frac{t}{T_{max} + betrand} \quad (12)$$

w = integer in the range [1,50], T = current number of iterations of the algorithm, T_{max} = maximum number of iterations set for the algorithm, betarnd = random number generated by the [0, 1] beta distribution, The μ's and γ's range are (0, 1).

10: Iterate from step 2 until the maximum number of cycles or accuracy requirements are fulfilled.

11: Exit the loop and output the algorithm results.

The features which reflect the device characteristics and behaviour related to various attacks are only considered. Each feature has been assigned a weight value.

Phase 3: Classification

For classifying the data as normal or anomaly, the XGBoost classifier is applied. eXtreme Gradient Boosting (XGBoost) is a boosting technique that belongs to the ensemble-based method. It includes constructing a series of decision trees, which is called as a sequential ensemble method. This method produces results with low bias and high variance, since the model has a strong capability to fit the training information.

EXPERIMENTAL RESULTS

The proposed optimized feature selection technique for DIDS has been implemented in Python 3.0 with Google Colab environment.

Dataset Description and Visualization

The KDD cup dataset contains the following 42 features with 494021 records. The format of the KDD cup dataset is shown in Figure 1.

	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot	num_failed_logins	logged_in	num_compromised	root_shell
0	0	tcp	http	SF	181	5450	0	0	0	0	0	1	0	(
1	0	tcp	http	SF	239	486	0	0	0	0	0	1	0	(
2	0	tcp	http	SF	235	1337	0	0	0	0	0	1	0	(
3	0	tcp	http	SF	219	1337	0	0	0	0	0	1	0	(
4	0	tcp	http	SF	217	2032	0	0	0	0	0	1	0	(

Figure 1 Dataset format

Classification Results

The performance of the optimized XGBoost classifier has been compared with the XGBoost with PCA and normal XGBoost classifier without applying any dimension reduction or optimization techniques. The classification performance is evaluated in terms of the following measures : Accuracy and F1-score

Table 1 and Figure 2 show the comparison results of accuracy and F1-score for these 3 approaches.

Techniques	Accuracy	F1-score
XGBoost-Optimized	99.65	98.35
XGBoost-Normal	97.35	96.4
XGBoost-PCA	98.6	97.58

Table 1 Comparison results of accuracy and F1-score

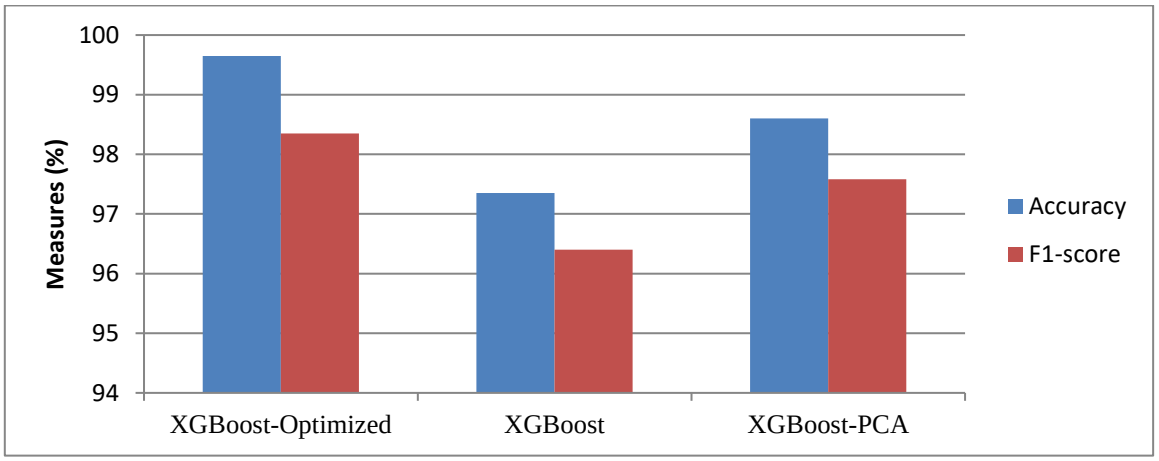

Figure 2 Comparison Results

As seen from Figure 2, the optimized XGBoost classifier attains highest accuracy of 99.65% and highest F1-score of 98.35% , when compared to the other two classifiers.

CONCLUSION

In this paper, an optimized feature selection technique for Distributed Intrusion Detection System (DIDS) in IoT environments has been proposed. During the preprocessing phase, t-SNE technique is applied for data exploration and visualizing the high-dimensional data and PCA technique is applied for dimensionality reduction. Then, for selecting the optimal features from the preprocessed dataset, the IGSA is applied. Finally, for classifying the data as normal or anomaly, the XGBoost classifier is applied. In this work, the KDD cup dataset is used for training, which contains the 42 features with 494021 records. The performance of the optimized XGBoost classifier has been compared with the XGBoost with PCA and normal XGBoost classifier without applying any dimension reduction or optimization techniques. Experimental results show that the optimized XGBoost classifier attains highest accuracy of 99.65% and highest F1-score of 98.35% , when compared to the other two classifiers.

REFERENCES

[1] Madhu, B., Chari, M. V. G., Vankdothu, R., Silivery, A. K., &Aerranagula, V. (2023). Intrusion detection models for IOT networks via deep learning approaches. *Measurement Sensors*, *25*, 100641. https://doi.org/10.1016/j.measen.2022.100641

[2] Facchini, S., Giorgi, G., Saracino, A., & Dini, G. (2020). Multi-level Distributed Intrusion Detection System for an IoT based Smart Home Environment. *6th International Conference on Information Systems Security and Privacy*. https://doi.org/10.5220/0009170807050712

[3] Alsakran, F., Bendiab, G., Shiaeles, S., & Kolokotronis, N. (2020). Intrusion Detection Systems for Smart home IoT Devices: Experimental Comparison study. In *Communications in computer and information science* (pp. 87–98). https://doi.org/10.1007/978-981-15-4825-3_7

[4] Vijayan PM, Sundar S (2023) An automated system of intrusion detection by IoTaided MQTT using improved heuristic-aided autoencoder and LSTM-based Deep Belief Network. PLoS ONE 18(10): e0291872. https://doi.org/10.1371/journal.pone.0291872

[5] Poongodi, M., & Hamdi, M. (2023). Intrusion detection system using distributed multilevel discriminator in GAN for IoT system. *Transactions on Emerging Telecommunications Technologies*, *34*(11). https://doi.org/10.1002/ett.4815

[6] Otoum, Y., Liu, D., & Nayak, A. (2019). DL-IDS: a deep learning–based intrusion detection framework for securing IoT. *Transactions on Emerging Telecommunications Technologies*, *33*(3). https://doi.org/10.1002/ett.3803

[7] Bakhsh, S. T., Alghamdi, S., Alsemmeari, R. A., & Hassan, S. R. (2019). An adaptive intrusion detection and prevention system for Internet of Things. *International Journal of Distributed Sensor Networks*, *15*(11), 155014771988810. https://doi.org/10.1177/1550147719888109

[8] Qaddoura, R.; Al-Zoubi, A.M.; Almomani, I.; Faris, H. A Multi-Stage Classification Approach for IoT Intrusion Detection Based on Clustering with Oversampling. Appl. Sci. 2021, 11, 3022. https://doi.org/10.3390/app11073022

[9] Sohail, S., Fan, Z., Gu, X., & Sabrina, F. (2022). Multi-tiered Artificial Neural Networks model for intrusion detection in smart homes. *Intelligent Systems With Applications*, *16*, 200152. https://doi.org/10.1016/j.iswa.2022.200152

[10] Wang, M.; Yang, N.; Weng, N. Securing a Smart Home with a Transformer-Based IoT Intrusion Detection System. Electronics 2023, 12, 2100. https://doi.org/10.3390/ electronics12092100

[11] 11.http://kdd.ics.uci.edu/databases/kddcup99/kddcup.data_10_percent.gz

[12] Moses Njue and Billy Franklin,"Dimensionality Reduction on MNIST dataset using PCA, T-SNE and UMAP",2020

[13] 13.Yang, Z.; Cai, Y.; Li, G. "Improved Gravitational Search Algorithm Based on Adaptive Strategies", Entropy 2022, 24, 1826.https://doi.org/10.3390/e24121826