LAUTECH Journal of Engineering and Technology 18 (2) 2024: 148-160

10.36108/lauiet/4202.81.0241

Development of an intrusion detection system using Mayfly feature selection and artificial neural network algorithms

[1*]Abdulsalam S. O., [1]Ayofe R. A., [2]Edafeajiroke M. F., [1]Ajao J. F. and [1]Babatunde R. S.

[1]*Department of Computer Science, Faculty of Information and Communication Technology, Kwara State University, Malete, Nigeria.*

[2]*Department of Mathematics and Computer Science, Faculty of Natural and Applied Sciences, Michael and Cecilia Ibru University Agbarha-Otor Delta State, Nigeria.*

Article Info

Article history:

Received: May 6, 2024
Revised: June 20, 2024
Accepted: June 24, 2024

Keywords:

Mayfly Algorithm (MA), Intrusion Detection System, Artificial Neural Network, Dimensionality Reduction.

Corresponding Author:
sulaiman.abdulsalam@kwasu.edu.ng

ABSTRACT

Protecting the privacy and confidentiality of information and devices in computer networks requires reliable methods of detecting intrusion. However, effective intrusion detection is made more difficult by the enormous dimensions of data available in computer networks. To boost intrusion detection classification performance in computer networks, this study proposed a feature selection mode for the classification task. The proposed model utilized the Mayfly feature selection algorithm and ANN as the classifier. The model was also tested without a Mayfly Algorithm (MA). The efficiency of the model was determined through a comparison of its Accuracy, Specificity, Precision, Negative predictive value, False positive rate, False discovery rate, False negative rate, Sensitivity, and F1- score. The experimental outcomes revealed that the proposed model is more efficient than existing models when implemented on the Canadian Institute for Cybersecurity Intrusion Detection System 2017 (CIC-IDS 2017) dataset. Accuracy of 99.94% (using Data+mayfly+ANN) and 90.17% (using Data+ANN) were attained after experimentation. The proposed model demonstrated superior accuracy compared to existing studies. Its robustness is due to employing mayfly techniques that combine the strengths of PSO, GA, and FA for optimal feature selection. This research presents a dependable dimensionality reduction model useful for intrusion detection and improving security in computer networks.

INTRODUCTION

Presently, the Internet can be considered to be an undeniable part of people's daily lives, and the total number of internet users has grown, exponentially. As a result, these users are eager to transmit a higher volume of critical data through the wires (Chatterjee *et al*., 2023). Therefore, the infrastructure built for data transmission should consider security issues to create a reliable, accurate, and configurable security system (Li *et al.*, 2022). To protect the security and integrity of user data, various tools, such as firewalls, antivirus, encryption, and authentication applications are in place. However, the mentioned tools have not been efficient enough to safeguard the systems against various types of threats (Alagrash, 2023).

Moreover, when it comes to the capability of detecting attacks, these tools face difficulty in separating Denial of Service (DoS) attacks from normal traffic (Bhattacharyya et al., 2020). To improve the security of the networks, it is suggested to combine firewalls with intrusion-detection systems (Zervoudakis and Tsafarakis, 2020). Saheed *et al*. (2022) describe the intrusion detection process as a series of steps enabling the monitoring, detection, and analysis of activities that violate network security policies. Li *et al.* (2022) proposed a framework for detecting network attacks, this framework is based on the assumption that security violations can be identified by monitoring system audit records for abnormal patterns of system usage.

In other words, the attacker's behaviour can be considered to be a basis for anomaly-based detection systems (Choudhary and Kesswani, 2021). These systems define malicious behaviour as an activity that demonstrates a deviation from the regular operation. The most significant benefit that these systems can provide for the users is the influential detection rate of both known and unknown attacks. A properly tuned IDS guarantees deeper insight into networks by providing visibility and control measures to minimize threats and attacks. Machine-learning algorithms can bring a lot of advantages for the daily monitoring of the network systems, but these techniques should be adopted to increase the attack detection rate and to decrease the system complexity. As the data collected from networks are high-dimensional, processing this data in the format as they are collected, makes the system inefficient. The irrelevant features should be removed using proper feature selection to decline the computation time, increase prediction performance, and recognize the pattern involved in the developed methods (Ahmad *et al.,* 2022). In other to improve the current intrusion detection system. Hence the study. Machine Learning (ML) models have been widely employed in Intrusion Detection Systems (IDS) to enhance accuracy and effectiveness in computer networks. However, considering the reliability of current state-of-the-art models for practical applications in real-world computer networks is crucial (Maabreh *et al.,* 2022). While several evaluation metrics have shown promising results for ML models, the primary focus has been on achieving high performance on a simple dataset, ignoring practical applications (Li *et al*., 2021). The high cost of errors in IDS and the difficulty in collecting and storing all relevant features in a live computer network feed may have contributed to this trend. Additionally, the nature of ML allows for continuous improvement of its hyper-parameters, making it challenging to generalize its performance to different datasets (Ahmad *et al.*, 2022). Despite these challenges, there is a need for further research to enhance the reliability of ML-based IDS in computer networks.

In IDS, feature selection methods are one of the major dimensionality techniques that have been applied to reduce the dimensionality of the datasets. However, feature selection methods come with two conflicting objectives which are the minimization of the number of features and classification error. The single-objective feature selection approaches applied in IDS are not able to confront both objectives simultaneously reduction is a technique used to reduce the complexity of high-dimensional data and keep only the most important information (Jyothsna *et al.,* 2020). This can be achieved through various methods, such as feature selection, feature extraction, or feature compression. Dimension reduction can help improve the accuracy and computational efficiency of intrusion detection (Zhao *et al.,* 2017). The motivation for this research is to address the challenges faced by IDSs in handling high-dimensional data and to improve the accuracy of IDS in computer networks. This study proposes a dimensionality reduction technique based on a MA to fetch relevant information from the given dataset. The optimal feature subset from the MA was used to build an ANN classifier. ANN has been a good classifier but it is computationally intensive, interpretability issues, and selecting appropriate architecture remain debatable hence this study employs mayfly for an optimal performance of the ANN model. According to Chatterjee et al.

(2023), IDS are critical components in ensuring the security of computer networks. Zhang *et al.* (2017) found out that IDS model often experiences an increase in false positive and false negative rates due to high-dimensional data generated from high numbers of users on the networks. Furthermore, IDSs need to be able to handle large amounts of data in real-time to be effective in detecting intrusions. ANN has been known to be a good classifier but has high local optimal issues as well as the challenge of selecting appropriate architecture for a given task. When attack frequency reduces, ANN becomes susceptible. Low-frequency assaults have an insufficient learning sample size. Because ANN cannot easily learn these assaults' properties, detection accuracy is low. Despite the advancements in IDS, accuracy and stability remain a concern for most IDS systems. Few research works have employed dimensionality reduction techniques to resolve the challenge of high-dimensional data in IDS. However, outcomes in recent times have been unsatisfactory. This is due to computational overhead on the existing IDS (Ayesha and Hanif, 2020). Selecting an appropriate feature selection technique is crucial in ensuring that the most important information is preserved while minimizing the reduction in accuracy (Revathi and Ramesh, 2018). This study aimed to improve the accuracy and computational efficiency of intrusion detection systems. A recently developed MA was used to select the optimal features subset which in turn formed a basis for building ANN model for Intrusion detection. The mayfly harnesses the benefits of the Genetic Algorithm (GA), Particle Swam Optimization Algorithm (PSO), and Firefly Algorithm (FA) for improved performance.

Artificial Neural Networks (ANNs) are a popular ML algorithm that uses interconnected nodes to process input data and make predictions. While ANNs are inspired by biological neural networks in the brain, they are not a direct mimicry of human brains. ANNs can have multiple hidden layers, allowing them to learn complex patterns and relationships in the data, making them well-suited for nonlinear functions (Kilincer *et al.,* 2021). However, ANNs are not infallible, and factors such as the quality of training data, choice of hyperparameters, and the presence of confounding variables or noise can affect their performance. While training, ANNs can be computationally expensive, recent advancements in optimization algorithms and hardware acceleration have reduced training time and cost for deep learning models (Kilincer *et al.,* 2021). Backpropagation is a widely used and effective algorithm for optimizing the weights of ANNs through gradient descent, which can be extended to deep networks with many layers by using techniques like weight initialization, batch normalization, and skip connections (Amin and Kabir, 2022). Overall, ANNs are a flexible and powerful ML algorithm that can be used for a wide range of applications, but their success depends on careful tuning and attention to the underlying data and problem domain (Chua and Salam, 2022).

Mayfly Optimization Algorithm

Mayflies are insects that belong to the order Ephemeroptera, part of a group of insects known as Palaeoptera. These insects appear mainly during the month of May in the UK, thus having the name Mayfly. Immature mayflies spend several years growing as aquatic nymphs until they are ready to go to the surface as adult mayflies. Most male adults assemble in swarms a few meters above the water to attract the females. They perform a nuptial dance which involves characteristic up and down movement generating a pattern (Patil *et al.*, 2022). Female mayflies go to these swarms for mating. The mating process lasts only for a few seconds after which they drop the eggs in the water and the cycle continues (Bhattacharyya *et al.,* 2020). MA algorithm was developed by Zervoudakis and Tsafarakis (2020) and is a new method for solving feature selection problems. It is a hybrid method combining the advantages of classical optimization methods such as PSO, GA and FA. Li *et al*. (2022) showed that PSO needed modifications as it is likely to get stuck in a local optimum, especially for problems having a high dimension. The MA performs the necessary modifications, ther eby enabling the algorithm to have better performance across small and large-scale feature sets (Bhattacharyya *et al.,* 2020). pseudocode for the Mayfly optimization process is shown in the Algorithm 1.1.

Algorithm 1.1: Mayfly Optimization Algorithm Li *et al.* (2022)

```
Objective function f(x), x=(x_1,…,x_d )^T
Initialize the male mayfly population x_i (i=1,2,…,N) and velocities v_mi
Initialize the female mayfly population y_i (i=1,2,…,M) and velocities v_fi
Evaluate solutions
Find global best gbest
Do While stopping criteria are not met
    Update velocities and positions of males and females
    Evaluate solutions
    Rank the mayflies
    Mate the mayflies
    Evaluate offspring
    Separate offspring to male and female randomly
    Replace worst solutions with the best new ones
    Update pbest and gbest
end while
```

REVIEW OF RELATED WORKS

Chatterjee *et al.* (2023) developed a resilient framework for identifying known and unknown network attacks using anomaly and signature-based techniques. Tested on publicly available datasets, the framework achieved accuracies of 90.94% and 99.67%, respectively. However, this study fails to address attacks from Distributed Denial of Service Attacks (DDoS) attacks.

Alagrash (2023) performed a Systematic Literature Review (SLR) that evaluates the viability of this solution type and highlights future research directions. In addition, it should give details on the most prevalent methodologies, enabling the identification of the most prevalent ML algorithms, architectures, and datasets.

Deore and Bhosale (2022) focused on developing an intrusion detection system using the Dolphin Atom Search Optimization (DASO) algorithm. The challenge of ensuring fast and efficient pre-processing independent of research was addressed. Additionally, the issue of increasing zero-day attacks was tackled, emphasizing the need for security systems capable of accurately detecting previously unknown attacks. This approach involved utilizing active feature optimization methods to create a generic meta-heuristic scale. This scale was capable of detecting both known and undiscovered assaults with high detection rates and low false alarm rates.

Saheed *et al.* (2022) developed ML-IDS for IoT network intrusion detection using supervised machine learning algorithms. UNSW-NB15, containing various attack types and normal network traffic was used. Feature scaling using min-max normalization was done to prevent information leakage on the test data. Principal Component Analysis (PCA) was used to reduce the high-dimensional dataset. Six models were evaluated based on accuracy, AUC, recall, F1, precision, kappa, and Mathew correlation coefficient (MCC). The model Achieved a high accuracy of 99.9% and MCC of 99.97%, demonstrating competitive performance compared to existing methods. the need to address other types of attack hence the study.

Almiani et al., (2020) worked on ML techniques for intrusion inherent flow. to better understand the status of machine learning, Almiani et al. analyzed 49 related works from 2009 to 2014 that focused on single, hybrid, and ensemble classifier design architecture. result showed that the dataset in this area is huge hence the need for a dimensionality reduction technique for an improved detection model, hence the study.

Shenfield et al. (2018) present an approach for the detection of malicious network traffic using ANN. The proposed ANN architecture obtains an average accuracy of 98%, an average area under the receiver operator characteristic curve of 0.98, and an average false positive rate of less than 2% in repeated 10-fold cross-validation.

Zhao *et al.* (2017) applied PCA to reduce the dimensionality of network traffic on the Knowledge Discovery and Data Mining 1999 (KDD-99) dataset. the reduced features were used to train two ML algorithms Softmax Regression and k-Nearest Neighbors (KNN). KNN achieved higher accuracy 85.24% with 3 dimensions and 85.19% with 6 dimensions compared to Softmax Regression 84.999% with 3 dimensions, dropping to 84.4% with more features. The model effectively identified both normal and malicious traffic. The reduced dimensionality lowered computational complexity, making it suitable for real-time IDS in Internet of Things (IoT) environments. Zhao et al. suggested unsupervised learning algorithms for feature selection. Investigating the effectiveness against other types of attacks is crucial. Also Utilising memory-efficient algorithms for resource-constrained environments remains a future concern.

Yuansheng *et al.* (2019) proposed a real-time network intrusion detection system using deep learning, big data, and natural language processing to address issues like low accuracy, high false positive rates, and inability to handle new intrusion types. data cleaning, coding, extraction and integration, and normalization were the preprocessing techniques employed. The system uses Flume for log collection and AEAlexJNet for deep learning. model was experimented on the KDD 99 dataset. AE-AlexNet model achieved an accuracy of 94.32%.

Tan *et al.* (2016) used linear discriminant analysis (LDA) for dimension reduction in intrusion detection. The result showed that LDA improved the accuracy of the intrusion detection system compared to other dimensionality reduction techniques. Nagpal *et al.* (2021) proposed an optimization algorithm combined with SVM for IDS. The main idea is to utilize Big Bang Big Crunch Optimization to select some features from the KDD-99 datasets. 41 features, and then utilize SVM Machine to compute the accuracy of these features.

Buczak and Guven (2016) explore machine learning and deep learning methods for cyber security, focusing on misuse and anomaly detection, highlighting the challenges in establishing their effectiveness due to complexity and richness. It emphasizes the need for labelled datasets and proposes investigating fast incremental learning methods to update models daily as a potential research area.

METHODOLOGY

In this study, the system goes through two stages of development which are pre-processing and classification. Removal of Missing data and selection of relevant features were the data preprocessing activities done. The preprocessed dataset was then passed into the ANN classifier for classification purposes as described in Figure 1. In addition, the high-dimensional dataset was also passed into the ANN for performance comparison.

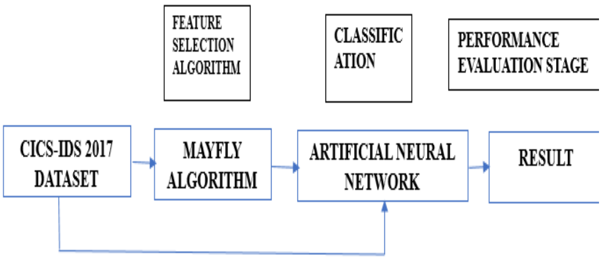

FEATURE SELECTION ALGORITHM
CLASSIFIC ATION
PERFORMANCE EVALUATION STAGE
CICS-IDS 2017 DATASET
MAYFLY ALGORITHM
ARTIFICIAL NEURAL NETWORK
RESULT

Figure 1: Framework of the Research Approach

Data Collection and Description

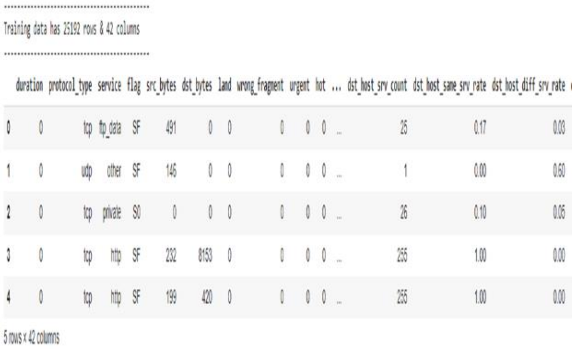
To evaluate the effectiveness of the dimension reduction system for intrusion detection, the CIC-IDS 2017 dataset from the Kaggle repository was used. This dataset includes both benign and malicious traffic, reflecting real-world scenarios. It was processed using the CICFlow Meter network traffic analysis tool, generating labelled flows based on parameters like timestamps, IP addresses, ports, protocols, and attack vectors. The processed data was stored in CSV files. The dataset, described in Table 1, contains five days of traffic data, with Thursday afternoon and Friday data suitable for binary classification. This study focused on the Friday data, which includes a DDoS attack, due to its rapid impact on systems and the significant volume of attacks from multiple sources to a single target. The dataset comprises 284,315 instances and 79 features.

During the data pre-processing stage it was observed that the raw dataset contains 65410 instances having missing class labels and 203 instances having missing information. By removing such missing instances, the dataset instance was reduced to 218702 instances. The data pre-processing involved the removal of missing data and then feature selection with the Mayfly algorithms.

Table 1: Description of files contained in CIC-IDS 2017 Dataset

Name of Files	Attacks Found
Monday-Working_hours	Benign (Normal human activities
Tuesday WorkingHours.pcap	Benign, FTP-Patator, SSH-Patator
WednesdayworkingHours.pcap ISCX.csv	Benign, DOS GoldenEye, Dos Hulk, DOS Slowhttptest, DOS slowloris, Heartbleed
Thursday-WorkingHoursMorning-WebAttacks.pcap_	Benign, Web Attack — Brute Force, Web Attack — Sql Injection, web Attack - XSS
Thursday-WorkingHoursAfternoon-Infilteration.pcap	Benign, Infiltration
Friday_WorkingHours_Mornin	Benign, Bot
Friday-WorkingHours-AfternoonPortScan	Benign, PortScan
Friday-WorkingHours AfternoonDDos	Benign, DD0S

The pre-processed data was then used as an input for building an ANN classifier model. data collection and preparation process ensured that the dataset used in this study accurately reflects real-world data and provides a suitable basis for evaluating the effectiveness of the proposed dimension reduction system for intrusion detection. After applying MA to select the optimal features the illustration of the training data is presented in Figure. 2.



```
----------------------------------------
Training data has 25192 rows & 42 columns
----------------------------------------
   duration protocol_type service flag src_bytes dst_bytes land wrong_fragment urgent hot ... dst_host_srv_count dst_host_same_srv_rate dst_host_diff_srv_rate
0         0           tcp ftp_data   SF       491         0    0              0      0   0 ...                 25                   0.17                   0.03
1         0           udp    other   SF       146         0    0              0      0   0 ...                  1                   0.00                   0.60
2         0           tcp  private   S0         0         0    0              0      0   0 ...                 26                   0.10                   0.05
3         0           tcp     http   SF       232      8153    0              0      0   0 ...                255                   1.00                   0.00
4         0           tcp     http   SF       199       420    0              0      0   0 ...                255                   1.00                   0.00
5 rows × 42 columns
```

Figure 2: Loaded Intrusion dataset

Intrusion Detection Benchmark Datasets

The Network Intrusion Detection System (NIDS) uses up-to-date data to accurately detect anonymous intrusions. KDD-99 dataset limits new attack types, while the CICID 2017 dataset provides a comprehensive representation of various attack types. The CIC Flow Meter analyzes network traffic, analyzing features like timestamps, protocols, and ports. Reliable datasets meet the required criteria.

CICIDS2017 Dataset

The CICID2017 dataset, based on abstracted network activity from 25 users, covers various attack scenarios like brute force, HeartBleed, botnet, DoS, DDoS, web, and infiltration attacks. It is available in two formats: PCAP format for packet payloads and CSV files for labelled flows, suitable for machine and deep learning application

NSL/KDD Dataset

The NSL-KDD dataset, a simplified version of KDD 99, evaluates intrusion detection system effectiveness with 41 features, 21 attacks in training and 37 in test. Hence it is pertinent to employ more recent and realistic datasets hence the choice of CICID2017 dataset.

Center for Applied Internet Data Analysis (CAIDA)(2002/2016)

The CAIDA dataset, a collection of anonymized data from San Jose's OC48 link, DDoS attack dataset, and 2016 Internet trace, has limitations that limit its effectiveness.

Feature Selection Process using Mayfly

During the development of the intrusion detection system, the first process used was to select the important features from the dataset by using the Mayfly feature selection techniques on the dataset. Before proceeding to the classification stage, the feature selection technique helped to ensure that the dimensionality of the selected data was reduced to avoid any form of under or overfitting during the classification stages. These two processes ensure efficient data processing while still preserving important information for intrusion detection.

ANN Classification

The second step used in this research is to develop the system and assess the performance of the ANN classification algorithms on the reduced dataset that was used to develop the system. These algorithms were used to classify the reduced data into intrusion categories. The performance of the algorithm was assessed based on metrics such as accuracy, specificity, sensitivity and precision.

Classification Performance Measurement

The final method used in this research is the assessment of the accuracy, specificity, sensitivity and precision. The formula for the matrices is shown in Table 2. The results of the proposed Mayfly and ANN intrusion detection system were compared with the existing intrusion detection systems.

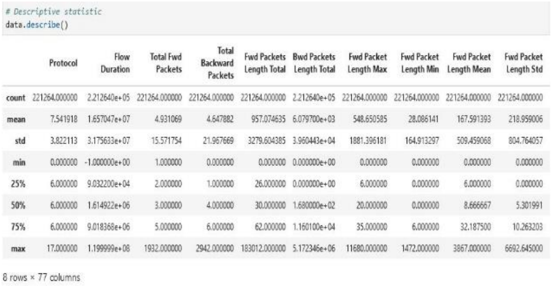
Table 2: Formula for Classification Performance Measurement

Measure	**Formula**
Precision	$\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$
Sensitivity	$\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$
Accuracy	$\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$
Specificity	$\frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$

RESULTS AND DISCUSSIONS

In this research, Python 3.7 executed in Jupyter Notebook was used for implementation. The experiments ran on a MacBook Pro with a 3.6GHz quad-core Intel Core i7 processor, 2GB GPU, and

16GB RAM. Data preprocessing involved removing missing information and using the MA for feature selection, with data scaled between 0 and 1. The dataset was split into 70% for training and 30% for testing, with the testing set further split for validation. Figure 3. Describe the train data after splitting. The experiments evaluated the system's Accuracy, Specificity, Precision, Negative predictive value, False positive rate, False discovery rate, False negative rate, Sensitivity, and F1- score using ANN classifier.



```
# Descriptive statistic
data.describe()
```

	Protocol	Flow Duration	Total Fwd Packets	Total Backward Packets	Fwd Packets Length Total	Bwd Packets Length Total	Fwd Packet Length Max	Fwd Packet Length Min	Fwd Packet Length Mean	Fwd Packet Length Std
count	221264.000000	2.212640e+05	221264.000000	221264.000000	221264.000000	2.212640e+05	221264.000000	221264.000000	221264.000000	221264.000000
mean	7.541918	1.657047e+07	4.931069	4.647882	957.074635	6.079700e+03	548.650585	28.086141	167.591393	218.959006
std	3.822113	3.175633e+07	15.571754	21.967669	3279.604385	3.960443e+04	1881.396181	164.913297	509.459068	804.764057
min	0.000000	-1.000000e+00	1.000000	0.000000	0.000000	0.000000e+00	0.000000	0.000000	0.000000	0.000000
25%	6.000000	9.032200e+04	2.000000	1.000000	26.000000	0.000000e+00	6.000000	0.000000	6.000000	0.000000
50%	6.000000	1.614922e+06	3.000000	4.000000	30.000000	1.680000e+02	20.000000	0.000000	8.666667	5.301991
75%	6.000000	9.018368e+06	5.000000	6.000000	62.000000	1.160100e+04	35.000000	6.000000	32.187500	10.263203
max	17.000000	1.199999e+08	1932.000000	2942.000000	183012.000000	5.172346e+06	11680.000000	1472.000000	3867.000000	6692.645000

8 rows × 77 columns

Figure 3: Description of the training data after splitting.

Result of Developed Intrusion Detection System

The intrusion detection system was developed using ANN, and the result of the classification achieved was visualized with training and validation Loss over Epochs as shown in Figure. 4. After the system has been developed, the details of each amount of fraudulent transaction against the normal transactions recorded are provided in Figure 5. These results indicate that the developed system was highly effective in detecting network intrusions while maintaining a low false positive rate and false discovery rate. The results provide empirical evidence of the effectiveness of the developed system for intrusion detection in computer networking. The result of the experiment shows that the system achieved an accuracy of 99.94%, specificity of 99.96% and precision of 99.97%. Monitoring the training and validation loss over epochs helps in understanding how well a model is learning and whether it is overfitting, underfitting, or performing optimally. By plotting the training and validation loss over epochs on the same graph, the training dynamics of the model as shown in Figure 4 and Table 3 was observed.

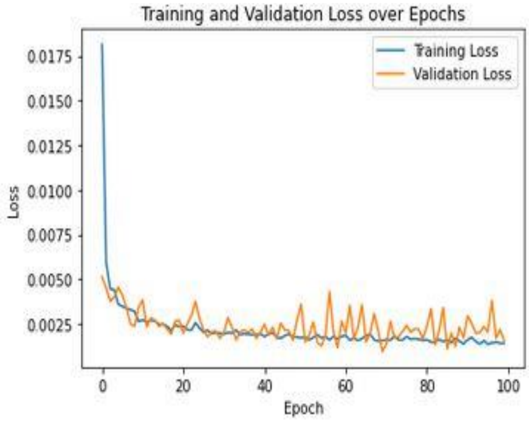
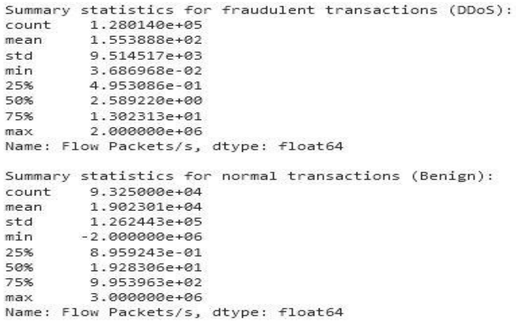


Figure 4: Training and Validation Loss over Epochs

Since both training and validation loss are decreasing over epochs, it indicates that the model is learning and generalizing well. This is a sign of good model training.



```
Summary statistics for fraudulent transactions (DDoS):
count     1.280140e+05
mean      1.553888e+02
std       9.514517e+03
min       3.686968e-02
25%       4.953086e-01
50%       2.589220e+00
75%       1.302313e+01
max       2.000000e+06
Name: Flow Packets/s, dtype: float64

Summary statistics for normal transactions (Benign):
count     9.325000e+04
mean      1.902301e+04
std       1.262443e+05
min      -2.000000e+06
25%       8.959243e-01
50%       1.928306e+01
75%       9.953963e+02
max       3.000000e+06
Name: Flow Packets/s, dtype: float64
```

Figure 5: Description of Fraudulent and normal transactions after classification

Comparative analysis

In this section, a comparison analysis of results was made with previous studies in the field. The accuracy metric was chosen as a measure of comparison due to its widespread use in evaluating model performance. Table 4 depicts a comparison of the accuracy metrics of the developed MA and ANN IDS with the results from other studies that employed ML algorithms for intrusion detection.

Table 3: Result of the experiment

Performance measures (%)	Data + ANN	Data + Mayfly + ANN	Formulae
Accuracy	98.90	99.94	$\frac{(TP+TN)}{(P+N)}$
Specificity	98.91	99.96	$TN/(FP+TN)$
Precision	99.15	99.97	$TP/(TP+FP)$
Negative predictive value	99.23	85.07	$TN/(TN+FN)$
False positive rate	0.15	7.19	$FP/(FP+TN)$
False discovery rate	0.16	0.58	$FP/(FP+TP)$
False negative rate	0.65	13.02	$FN/(FN+TP)$
Sensitivity	99.33	87.93	$TP/(TP+FN)$
F1- score	98.91	91.15	$2TP/(2TP+FP+FN)$

This comparison was limited to studies that used similar approaches to ensure that the criteria for comparison were aligned with the scope of this study. The results indicated that the model developed in this research showed a better accuracy performance compared to previous studies shown in Table 4.

Discussion and Interpretation of Results

The experiment results showed that MA was highly influential in detecting network intrusions while maintaining a low false positive rate and false discovery rate. The system's Accuracy, Specificity, Precision, Negative predictive value, False positive rate, False discovery rate, False negative rate, Sensitivity, and F1- scores were 99.94%, 99.96%, 99.97%, 85.07%, 7.19%, 0.58%, 13.02%, 87.93%, 91.15% respectively.

Table 4: Comparative analysis of results with related work

Authors	Methods	Accuracy
Rincy and Gupta (2021)	Decision trees	80.77 %
Shin *et al.*, 2020	Decision trees and random forests	88.89%
Saranya *et al.* (2020)	LDA algorithm and Random Forest	88.1%.
Wani *et al.* (2019)	SVM and CART algorithm	87.60%
Present Research	**Data + Mayfly + ANN**	**99.94%**

These results indicate that the proposed system effectively reduced the dimensionality of the data while preserving the relevant information for intrusion detection. The high accuracy, specificity, and precision demonstrate the system's ability to effectively distinguish between normal and malicious network traffic, which is essential for int rusion detection. Additionally, the low false positive rate and false discovery rate suggest that the system has a low rate of incorrect detections, which can reduce the risk of false alarms in network security. The results also showed that the proposed system achieved high processing efficiency by using mayfly techniques for feature selection and ANN for classification. The high processing efficiency of the proposed system makes it suitable for real-time intrusion detection in computer networking. It can be deduced that the results of the proposed dimension reduction system for intrusion detection showed that the system could provide better accuracy, specificity, sensitivity, and precision than other existing methods. The results of this study confirmed the effectiveness and efficiency of the proposed system for intrusion detection in computer networking.

CONCLUSIONS

The increasing sophistication of cyber-attacks has made intrusion detection critical. This research utilized MA for feature selection to reduce the dimensionality of the dataset, followed by classification using an Artificial Neural Network (ANN) to identify intrusions. The study evaluated the system's performance based on Accuracy, Specificity, Precision, Negative predictive value, False positive rate, False discovery rate, False negative rate, Sensitivity, and F1- score. Results were compared with other methods, focusing primarily on accuracy. The CIC-ID2017 dataset, sourced from Kaggle, was pre-processed and split into training (70%) and testing (30%) sets. The ANN classifiers were trained on both the full and reduced datasets, achieving 98.90% accuracy without feature selection and 99.94% with feature selection. The model was assessed using various performance metrics and compared to previous research. Results showed that the proposed system accurately detects common network attacks and is valuable for intrusion detection, especially with large data and limited computational resources. Future research should address validating performance in more complex networks, detecting new threats, and considering false alarms. the proposed dimension reduction system for intrusion detection in computer networks is effective and promising, contributing to ongoing efforts to enhance network security against cyber threats.

RECOMMENDATION

Intrusion detection is critical to computer network security, helping organizations protect their networks from cyber threats. Extending the study to other IDS, such as host-based and network-based systems, would be valuable. several avenues for future research in intrusion detection and dimension reduction are Integrating additional feature selection and extraction techniques, this is to enhance model performance. Incorporate other classification algorithms, such as Random Forest and Decision Tree, to further evaluate and improve the performance. Other techniques can be explored to further reduce the false positive rate, addressing a common challenge in intrusion detection and hence can lead to more effective solutions for intrusion detection in computer networks.

REFERENCES

Abbas, A., Khan, M. A., Latif, S., Ajaz, M., Shah, A. A., and Ahmad, J. (2022). A New Ensemble-Based Intrusion Detection System for Internet of Things. Arabian Journal for Science and Engineering, 47(2), 1805–1819. https://doi.org/10.1007/s13369-021-06086-5

Ahmad, K., Maabreh, M., Ghaly, M., Khan, K., Qadir, J., and Al-Fuqaha, A. (2022). Developing future human-centered smart cities: Critical analysis of smart city security, Data management, and Ethical challenges. Computer Science Review, 43,100452. https://doi.org/10.1016/j.cosrev.2021.100452

Alagrash, Y.H. Mehdy, H.S. Mahdi, R.H. (2023). A review of intrusion detection system methods and Techniques: past, present and future. International Journal on “Technical and Physical Problems of Engineering” (IJTPE), Iss. 54, Vol. 15, No. 1, Mar. 2023

Almiani, M., AbuGhazleh, A., Al-Rahayfeh, A., Atiewi, S., Abdul Razaque, and Razaque, A. (2020). Deep recurrent neural network for IoT intrusion detection system. Simulation Modelling Practice and Theory, 101, 102031. https://doi.org/10.1016/j.simpat.2019.102031.

Amin, Z., and Kabir, A. (2022). A Performance Analysis of Machine Learning Models for Attack Prediction using Different Feature Selection Techniques. Proceedings - 2022 IEEE/ACIS 7th International Conference on Big Data, Cloud Computing, and Data Science, BCD 2022, 130–135. https://doi.org/10.1109/BCD54882.2022.9900597

Ayesha, S., and Hanif, M. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. Elsevier. https://www.sciencedirect.com/science/article/pii/S156625351930377X

Bhattacharyya, T., Chatterjee, B., Singh, P. K., Lee, J. E., Geem, Z. W., and Sarkar, R. (2020). Mayfly in Harmony: A New Hybrid Meta-Heuristic Feature Selection Algorithm. IEEE Access, 8, 195929–195945. https://doi.org/10.1109/access.2020.3031718

Buczak, A. L., and Guven, E. (2016). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. IEEE Communications Surveys and Tutorials, 18(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

Chatterjee, S., Shaw, V., and Das, R. (2023). Multi-Stage Intrusion Detection System aided by Grey Wolf optimization algorithm. Research Square (Research Square). https://doi.org/10.21203/rs.3.rs- 2680915/v1

Choudhary, S., and Kesswani, N. (2021). A Hybrid Classification Approach for Intrusion Detection in IoT Network. In Journal of Scientific and Industrial Research (Vol. 80).

Chua, T.-H., and Salam, I. (2022). Evaluation of Machine Learning Algorithms in Network-Based Intrusion Detection System. http://arxiv.org/abs/2203.05232

Deore, B., and Bhosale, S. (2022). Intrusion Detection System with a Modified DASO Optimization Algorithm. International Journal on Engineering, Science and Technology, 4(1), 54–63. https://doi.org/10.46328/ijonest.66

Jyothsna, V., Sreedhar, A. N., Mukesh, D., and Ragini, A. (2020). A Network Intrusion Detection System with Hybrid Dimensionality Reduction and Neural Network Based Classifier. 187– 196. https://doi.org/10.1007/978-981-15-0936-0_19

Kilincer, I. F., Ertam, F., and Sengur, A. (2021). Machine learning methods for cyber security intrusion detection: Datasets and comparative study. Computer Networks, 188. https://doi.org/10.1016/J.COMNET.2021.107840

Li, M. Chen, H. Shi, X. Liu, S. Zhang, M. and Lu, S. (2022) Amulti-information fusion `triple variables with iteration' inertia weight PSO algorithm and its application," Appl. Soft Comput., vol. 84, Nov. 2019, Art. no. 105677. doi:10.1890/00129658(2002)083[0612:pcicim]2.0.co;2

Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, Y., Liu, X.,and He, B. (2021). A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/TKDE.2021.3124599

Maabreh, M., Obeidat, I., Elsoud, E. A., Alnajjar, A., Alzyoud, R., and Darwish, O. (2022). Towards Data-Driven Network Intrusion Detection Systems: Features Dimensionality Reduction and Machine Learning. International Journal of

Interactive Mobile Technologies, 16(14), 123–135. https://doi.org/10.3991/ijim.v16i14.30197

Nagpal, M., Kaushal, M., and Sharma, A. (2021). A Feature Reduced Intrusion Detection System with Optimized SVM Using Big Bang Big Crunch Optimization. Wireless Personal Communications, 1–27. https://doi.org/10.1007/s11277-021-08975-2

Revathi, M., and Ramesh, T. (2018). Network intrusion detection system using reduced dimensionality.

Rincy, T., and Gupta, R. (2021). Design and Development of an Efficient Network Intrusion Detection System Using Machine Learning Techniques. Wireless Communications and Mobile Computing, 2021,1–35. https://doi.org/10.1155/2021/9974270

Saheed, K. Y., Idris, A. A., Misra, S., Kristiansen, H. M., and Colomo-Palacios, R. (2022). A machine learning-based intrusion detection for detecting internet of things network attacks. Alexandria Engineering Journal, 61(12), 9395–9409 https://doi.org/10.1016/j.aej.2022.02.063

Saranya, T., Sridevi, S., Deisy, C., and Chung, T. (2020). Performance analysis of machine learning algorithms in intrusion detection system: A review. Elsevier Procedia Computer. https://www.sciencedirect.com/science/article/pii/S1877050920311121

Shin, Y., Advanced, K. K.-I. J. of, and 2020, undefined. (2020). Comparison of anomaly detection accuracy of host-based intrusion detection systems based on different machine learning algorithms. Pdfs.Semanticscholar.Org, 11(2).https://pdfs.semanticscholar.org/b53a/29af26ef8b24018dd3c2483444af494e3ad7.pdf

Tan, Z., Jamdagni, A., He, X., and Nanda, P. (2016). Network Intrusion Detection based on LDA for payload feature selection. 2016 IEEE Globecom Workshops,1545–1549. https://doi.org/10.1109/GLOCOMW.2010.5700198

Wani, A., Rana, Q., Saxena, U., and Pandey, N. (2019). Analysis and detection of DDoS attacks on cloud computing environment using machine learning techniques. Ieeexplore.Ieee. Org. https://ieeexplore.ieee.org/abstract/document/8701238/

Yuansheng, D., Wang, R., and He, J. (2019). Real-Time Network Intrusion Detection System Based on Deep Learning. 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS). https://doi.org/10.1109/icsess47205.2019.9040718

Zervoudakis, K. and Tsafarakis, S. (2020)``A mayy optimization algorithm," Comput. Ind. Eng., vol. 145, Jul. 2020, Art. no. 106559, doi:10.1016/j.cie.2020.106559

Zeyuan F. (2022). Computer Network Intrusion Anomaly Detection with Recurrent Neural Network. Mobile Information Systems, 2022, 1–11. https://doi.org/10.1155/2022/6576023

Zhang, X., Ding, S., and Xue, Y. (2017). An improved multiple birth support vector machine for pattern classification. Neurocomputing, 225, 119–128. https://doi.org/10.1016/j.neucom.2016.11.006

Zhao, S., Li, W., Zia, T., and Zomaya, A. Y. (2017). A Dimension Reduction Model and Classifier for Anomaly-Based Intrusion Detection in Internet of Things. 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data

Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 836–843. https://doi.org/10.1109/DASC-PICom DataCom-CyberSciTec.2017.141

Zhao, S., Li, W., Zia, T., and Zomaya, A. Y. (2017). A Dimension Reduction Model and Classifier for Anomaly-Based Intrusion Detection in Internet of Things. 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 836–843. https://doi.org/10.1109/DASC-PICom-DataCom-CyberSciTec.2017.14

Patil, R., Tamane, S., Rawandale, S. A., and Patil, K. (2022). A modified mayfly-SVM approach for early detection of type 2 diabetes mellitus. International Journal of Power Electronics and Drive Systems/International Journal of Electrical and Computer Engineering, 12(1), 524. https://doi.org/10.11591/ijece.v12i1.pp524-533.