International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 8 Aug 2023, pp: 633-644 www.ijaem.net ISSN: 2395-5252

Cyber security mechanism based on anomaly event detection

[1]Haider K. Hoomod, [2]Jolan Rokan Naif, [3] Qutaiba Humadi Mohammed

1: Mustansiriyah University, College of Education, computer science depart., Baghdad, Iraq,
2: Iraqi Commission for Computers & Informatics/Informatics Institute for Postgraduate Studies/Baghdad, Iraq
3: College of Nursing, University of Baghdad, Baghdad/Iraq

Date of Submission: 20-08-2023 Date of Acceptance: 31-08-2023

ABSTRACT
The Internet of Things (IoT) aims to create a smarter environment through constant connectivity and diversity. The primary asset and danger of this technology are crucial data processing duties across many businesses. Therefore, developing an efficient security system is its top concern. Data is captured and transmitted in real-time, making IoT ideal for the sensitive data concept. The consequences of a cyberattack on IoT vary owing to many factors. While well-documented in the data science literature, cybersecurity approaches the obstacle from an external perspective, focusing on network activity data rather than the target. Additionally, current solutions lack a data science strategy that aligns with needs by analyzing the threats' taxonomy, model, and environmental structure. Current frameworks focus on signature-based solutions, which are more reliable while not detecting zero-day threats. Successful approaches are based on thoughtful categorization and contextualization of the job at hand. This study focuses on evaluating cybersecurity history to identify potential cyber incidents, followed by discussing machine learning approaches optimized by swarm optimization algorithms to meet the required requirements. To provide a trustworthy and effective intrusion system with machine learning capabilities, a real-time streaming anomaly detection model was optimized to assess real-world vulnerabilities that might threaten daily life. This paper addresses using 11 classifiers from machine learning algorithm, also they optimized by proposed hybrid swarm optimization algorithm (ant colony, Gray Wolves, and PSO) the need for an efficient and reliable intrusion detection system (IDS) in ongoing research. Our real-time anomaly detection technology can identify zero-day attacks using cybersecurity backdrop analysis, a revolutionary data gathering and processing tool, model selection, and adaptation to the environment.
Keywords: Internet of Things, data processing, cybersecurity, real-time, zero-day, machine learning, intrusion detection systems, vulnerabilities, streaming, anomaly detection.

I. INTRODUCTION

On the battlegrounds of cyberspace, attackers frequently hide in the shadows but are always ready to pounce on their victims the moment they see an opening. The massive collecting of vast volumes of data from a variety of perspectives is nothing novel in today's modern world[1,2]. This may be extremely helpful for security safeguards but at the same time, it presents security operations teams with issues that have never been seen before.[3] Every day, the people who work in security operations are up breathing down their necks in large alerts, and they have to keep themselves occupied by assessing alerts, connecting alerts with occurrences, and attributing assaults based on their knowledge and skills.[4,5] Finding a way to analyze attackers from many dimensions and analyze their potentials before presenting the assessment findings to security operations people, who will then identify the most dangerous attackers, is an important task that has to be completed as soon as possible in order to handle these issues with security operations. The approach known as "attributed graph modeling" is a powerful tool that enables the modeling of attacks by drawing on components of their structures, attributes, and the passage of time[4,5].

The use of attributed networks is widespread. Examples of networks include social networking sites, networks for communication, and product co-purchase networks. These types of networks are defined by an architecture of nodes,

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 8 Aug 2023, pp: 633-644 www.ijaem.net ISSN: 2395-5252

with each node having specific qualities that are utilized to determine the features of the network[6]. In the field of cybersecurity, attributed diagrams are superior to other modeling approaches when it comes to data from cyberspace[7]. For instance, attacker characterization is the process of constructing a relational graph in cyberspace, with the perpetrators of the assault serving as nodes and the features of the attack serving as attributes[8]. An example of the application of attributed graph-based detection of anomalies in cybersecurity is the identification of attackers and the attack pathways they take inside large-scale cyberspace relational graphs.

Before utilizing an attributed graph for the purpose of attacker modeling, it is necessary for us to specify the graph's characteristics, vertices, and edges. In the graph that is being attributed, the vertices represent the attackers (or the IP addresses of the attackers) and the victims (or the IP addresses of the victims). If device logs are incorporated into the graph, then the graph will also contain representations of processes, files, and services[6,9,10].

An adjacency matrix is a matrix with a square shape that is used to show whether or not vertices in a graph are next to one another. This allows us to utilize the matrix to describe the structure of the graph. In terms of attributed graphs, the adjacency matrix illustrates the attacks that perpetrators of crimes have carried out against victims[10]. Vertex characteristics are properties that are associated with an attacker. These attributes include the IP address, whether or not the IP address is publicly accessible, the portion of the IP address, and the ports[10]. The edges represent attacks that were carried out by the assailants on the victims. Finding nodes in the behavior pattern that are quite different from the other nodes is the first step in implementing an attribute graph-based anomaly detection system. This detection approach may be quite helpful for identifying actual attackers and the behavior that is associated with them when it comes to the detection of network intrusions[11].

IDS technology differs in the events observed and the analytic methods used for identifying cyber threats. The two most frequent types of IDS are host-based (HIDS) and network-based (NIDS). NIDS tracks network traffic and application protocol activity across segments. HIDS monitors host or device characteristics and events, including HMI, SCADA servers, and operator/engineering workstations. Intrusion detection methods include signature-based, anomaly-based, and stateful protocol analysis. Signature-based detection uses string comparison to match known threat signatures to observed events. Multi-event assaults and unknown threats are ineffectual.[12,13]

Anomaly-based approaches can detect unknown assaults by comparing the normal activity profile to actual events, utilizing statistics, expert knowledge, and machine learning to find significant deviations[14]. Poorly capturing the complexities of industrial operations might lead to numerous false positives. Stateful protocol analysis detects deviations by comparing prepared patterns according to network protocol standards to observed behavior[15]. The study needs frequent updates as protocol standard specifications change.

Anomaly detection using attributed graphs is now a hot issue in the field of security research and has been increasingly prevalent in recent years[16]. The detection of anomalous nodes is a primary focus of some study, which is accomplished by first clustering the nodes and then examining the edge weights of various communities[17]. A number of research construct their anomaly analyses on the basis of the subspace grouping of node properties. Anomaly identification based on residual analysis is a topic that has been the focus[18,19]. This technique makes advantage of codec mistakes to determine the degree to which vertices or edges in a graph depart from the norm. The sections that follow provide an explanation of various common approaches to anomaly identification that are based on attributed graphs.[20]

II. CYBER SECURITY

Although the complex communication system has many benefits, such as improved energy efficiency, reliability, and manageability, it also makes the system more susceptible to cyberattacks because of the large number of devices and access points that function outside the conventional administrative domain[21].

Investigation into the consequences of cyberattacks on electrical systems is of the utmost importance given the potential for the power grid to fail, which might result in catastrophic occurrences.[22,23]

The primary cause of the blackouts in North America was a lack of system awareness. This highlights the need of conducting cyber-attack analysis in order to ensure a stable and dependable functioning of the power supply[24,25]. It is possible for a cyberattack to cause overload, which may destroy the equipment, or a fraudulent demand request, which could end up in a significant amount of energy being created. In addition, a malicious assault has the potential to bring about false

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 8 Aug 2023, pp: 633-644 www.ijaem.net ISSN: 2395-5252

negatives, often known as a fake overload state in a power system.[26,27]

There is also the possibility of more interruptions in other sectors of the smart grid and the infrastructure for electric vehicles. It is demonstrated in [40] and [41] which malicious attacks that include restricting communications with a device can result in the termination of services provided by computers located in substations.

Therefore, the detection of cyberattacks in real time is of the utmost importance for ensuring the dependability of the functioning of important infrastructure, such as smart grids. Monitoring of the system must take place online and in a continuous manner if it is to fulfill the requirements for detecting targeted cyberattacks and achieving attack resilience [28,29,30].

III. ANOMALY DETECTION

Hawkins [31] defined an anomaly as an observation that "deviates so much from other findings as to arouse suspicions that it was produced by a different mechanism" in 1980. Analyzing anomalies can help detect credit card transactions, follow network traffic patterns that may indicate a cyber-attack, and discover malignant tumors in MRI imaging.

An IDS, independent of detection architecture, assumes significant differences in intruder behavior from normal patterns, enabling detection of unauthorized activity. Thus, these systems identify anomalies. This area identifies hidden patterns useful for diagnosing malignant situations in several industries, including security, medical, finance, energy, and agriculture[32].

An anomaly can indicate a negative change, such as increased CPU usage, indicating DoS attacks, or a positive abnormal behavior, such as increased online product purchases. It is crucial to identify the usual necessities and their causes. Currently, there is no viable answer due to the difficulty in providing an efficient suggestion, making specialists hesitant to utilize these approaches[33].

For a successful practical anomaly detection algorithm in an IoT context, the following aspects must be covered by the online concept where must be learned as soon as it occurs on the sensitive data stream while using an actual word, IoT anomaly detection algorithm. To build an online solution, each item must be processed once during training and include as little computing complexity as feasible [34].

The Internet of Things, like any other technology in today's world, requires data in order to provide improved services to its consumers and to improve its overall performance. Creating settings that are more intelligent can make our lives easier by allowing us to save time, money, and energy. Therefore, gathering and analyzing data produces knowledge that can be put into action. This not only enhances decision-making but also assigns the efficacy, effectiveness, and productivity of each characteristic of every platform [35].

The processing of data from the Internet of Things is hampered by a lack of computing, network, and storage resources. These properties, which are aligned with the criteria that have been addressed, such as interoperability and heterogeneity, solve the key issues that IoT analytics provide[36].

Because of the features of the Internet of Things, the data that are gathered have a nature that is spread and in real time, as well as a high volume, a quick velocity, and diversity. In order to be able to give improved insights and decision making, it need processing that is efficient with regard to cost. The Internet of Things (IoT) presents a number of issues, one of which is the fact that its fast speed, along with the possibility of inadequate quality and interpretation, compromises the consistency and trustworthiness of the models [37,38].

Because we have to be capable to analyze and predict in actual time with a limited amount of memory and time resources, we must be ready to deal with the scenario in which the data distribution modifications periodically when fast big data satisfies the real world. As a result, the task of monitoring systems, whether it is addressing network flow or host performance metrics, may be viewed as a continuous stream of inputs, signifying data flowing in and out in a continuous fashion. This is true whether the task is addressing network flow or host performance measurements. A flow X is a sequence of N-dimensional examples, denoted by the notation xt, that has the potential to expand to infinity [39].

$X = \{x_1, x_2, .., x_t, x_{x+1}...\}$

In contrast, batch processing involves the complete storage and training of datasets, which might occur several times. As a consequence of this, the amount of time required for an output to become available is longer, and if the total number of instances continues to grow, the technique will no longer be able to successfully complete the work at hand [40].

The issue of anomaly detection will be discussed in the following subjects, along with the most significant concerns and approaches that are now available. In conclusion, we will discuss the techniques of machine learning in terms of the

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 8 Aug 2023, pp: 633-644 www.ijaem.net ISSN: 2395-5252

benefits and drawbacks they offer with relation to the identification of anomalies in an Internet of Things streaming environment. These are the issues that need to be taken into consideration in order to get the optimal model that is suitable for the task at hand and the data that is involved[41,42].

The precise needs and context of IoT data make selecting the suitable machine learning technique challenging. To construct a reliable intrusion detection system in IoT, the keywords in Figure 1 are the major restrictions for a true and successful cybersecurity system. [38,43]

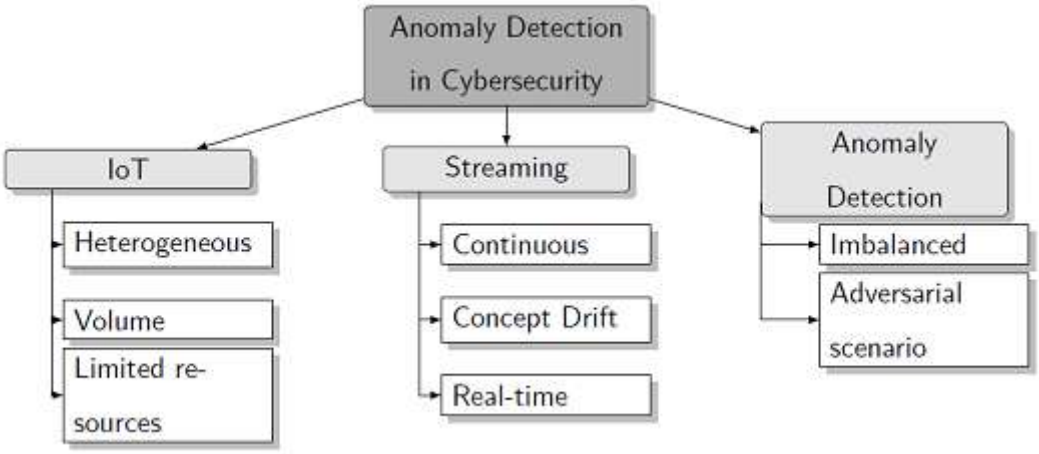

FIGURE 1: Domain contextualization

In an industrial network, network switches link network devices and segments, while firewalls at segment edges filter allowed traffic[44]. Network monitoring analyzes sections of the network (network traffic) and elements (switch and firewall records) to offer network transparency on participants, their networking properties, and communication linkages[45]. This information underlies anomaly detection systems. To identify abnormalities, it learns the network's architecture, communication linkages, time-related behavior, and communication content[46]. In its cybersecurity guidelines for production networks, Germany's Federal Office for Information Security (BSI) listed industrial network abnormalities and anomaly detection system feature requirements. The following features are needed to detect industrial network anomalies:

(1) Category A: general requirements
• overview of all devices communicating in the network,
• identification of protocols used in the network,
• identification of all the communication links in the network,

(2) Category B: unusual or exceptional activities in an ICS network
• identification of new devices in the network,
• identification of communications between two devices that was previously non-existent,
• identification of new protocols or changes in protocol among individual components,

(3) Category C: abnormal events in logs typical of production environments:
• identification of ICS-specific function codes that have not previously been used,
• the ability to determine whether an access attempt (e.g. read/write) pertains to an address that is not normally used by the device at hand,

(4) Category D: unusual changes in process data (sensor data, control data, etc.)
• identification of changes in time-related behavior,
• identification of deviations within defined value ranges

The BSI distinguishes between passive and active network monitoring and anomaly detection technologies. The passive data collection system uses network or wire taps without affecting network data or time-sensitive behavior. An active system generates data in industrial network traffic by sending queries to network gateways or devices. The system must be designed to disregard its own data for analysis. Additionally, consider time-sensitive behavior changes from increased data transfer while adopting this system.

IV. PROPOSED ANOMALY DETECTION SYSTEM

Misuse detection systems are unable to detect innovative network attacks since the attack signature doesn't exist in the database. Anomaly detection systems can identify new threats and alert the network before they do significant harm.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 8 Aug 2023, pp: 633-644 www.ijaem.net ISSN: 2395-5252

Anomaly detection, like abuse detection, requires a clear border between regular and aberrant traffic. The usual behavior profile is believed to differ considerably from the abnormal behavior profile. The profile of normal events and traffic should have a well-defined normal behavior. The normal behavior definition must include a computer machine's IP address or hostname, its VLAN, and the ability to sensitively track the target environment's typical behavior.

The normal profile should also include (i) occurrence patterns of particular system declarations in the application layer of the communications protocol stack; (ii) cooperation of data payloads with various components in application protocols; (iii) link patterns among secure servers and the Internet; and (iv) rate and burst length variations of every traffic type [13]. Network profiles must be flexible and self-learning from complicated and challenging traffic on the network to maintain accuracy and a small false approval rate.

Detecting malicious and abnormal traffic in a big data network is difficult and crucial. A massive amount of network data with a high-dimensional feature space is hard to evaluate and monitor. Monitoring and analyzing network traffic data requires effective data processing and pattern-learning techniques. Additionally, anomalous network traffic behaves similarly. In big-volume network traffic data, harmful and anomalous traffic of the exact same type is likely to happen repeatedly, although the number of occurrences is significantly lower than in regular data. The network traffic statistics are significantly unbalanced. Determining a normal zone or the boundary between normal and anomalous traffic is challenging, if not impossible. Different application areas define anomalies differently, complicating the situation. Labeled anomalous data for training and validation is often unavailable. Training and testing data include unknown distribution noise and normal and atypical behavior changes. All of these difficulties make network anomaly identification challenging.

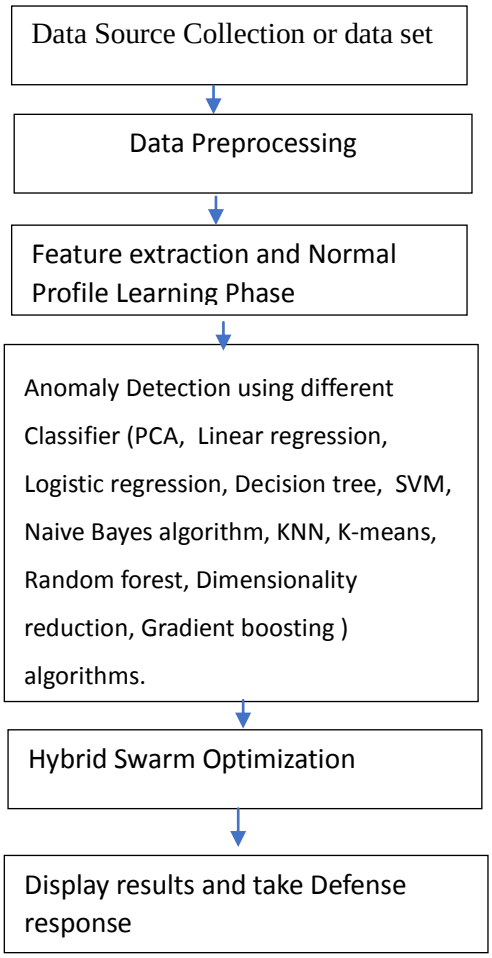

Figure 3. The Proposed modules Steps in an anomaly detection system.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 8 Aug 2023, pp: 633-644 www.ijaem.net ISSN: 2395-5252

Machine learning algorithms are useful for creating typical profiles and designing anomaly-based intrusion detection systems. For anomaly detection, normal network traffic data is generally available for model training. Most applications lack anomalous traffic-tagged data. We know that supervised machine learning algorithms require attack-free training data, we proposed to used the (PCA, Linear regression, Logistic regression, Decision tree, SVM, Naive Bayes algorithm, KNN, K-means, Random forest, Dimensionality reduction, Gradient boosting) algorithms. Thus, supervised learning requires labeled network data for regular and attack traffic. The machine learing algorithms used in this proposed system was optimized by using hybrid swarm optimization algorithm (HSOA) to get more efficient learning scutation and best results. HSOA is content a combination from three warm optimization (ant colony, Gray wolves, and PSO). However, prelabeled training data for each class is rare in real life. Most networks lack prelabeled training data and have significantly skewed traffic data. Few assault traffic records are blended with the majority of regular traffic records. The change in the network environment also affects regular traffic patterns, compounding the problem. Supervised intrusion detection systems (IDSs) generally have high false-positive rates (FPRs) due to the large differences between training and test datasets. Unsupervised learning methods used by anomaly detection systems are able to create a regular network traffic profile and system state to solve this challenge. Abnormal network activity is indicated by any divergence from normal. Therefore, semi-supervised and unsupervised machine learning approaches are often used in different security applications.

Misuse detection rules show the association between attribute conditions and class labels. Anomaly detection rules describe normal profiles of people, applications, and system services, as well as additional computer and network infrastructure resources. If it finds inconsistencies between program and user activity and system rules, an anomaly detection system should warn of a potential attack. An anomaly detection system needs extensive rules to operate.

Anomaly-based intrusion detection systems often employ associative categorization and association rules. Several theories have used association rules to create anomaly detection models. Association-rule anomaly detection systems typically take two phases. First, effective data mining is performed on system and network auditing records to uncover consistent and meaningful program and user behavior patterns. In the second stage, reliable classifiers are inductively developed utilizing the training dataset on important pattern characteristics to detect system or network traffic anomalies. In this proposed system, the NSL-KDD to training the classifiers. During the detection system's monitoring phase, event sequences that break the criteria are identified as cyberinfrastructure anomalies.

V. HYBRID SWARM OPTIMIZATION ALGORITHM (HSOA)

Even if the global best particle becomes locked in the local optimum, all particles in the initial PSO learn from it to update their location and velocity until the state of termination is achieved. This learning approach improves exploitation and offers quick convergence, but it fails with complicated search space issues. Now, ant colonies and gray wolves are offered as exemplary variations of PSO that restrict particle learning objects to maintain population variety. These tactics maximize exploration performance but slow convergence. To balance exploitation and exploration.In HSOA, tolerance-based searching direction modification is used. The process can cause the swarm to modify the search path to avoid local optimums and shorten the search space. We also use Gray Wolves learning to produce a candidate particle as a learning object for the swarm and parallel search on ant colony processing to achieve the goal of the swarm exploring different areas of the search space. To ensure the algorithm's efficiency and accuracy, we apply a prospective prediction technique to estimate the candidate particle's capacity to lead swarm exploitation in several dimensions utilizing Gray Wolves processing.

VI. PROPOSED OPTIMIZED MACHINE LEARNING IN HYBRID DETECTION

Misuse detection systems typically have a high rate of detection and low rates of false alarms because they match attack signatures with network events. However, existing systems cannot identify new assaults. However, the detection of anomalies helps systems define network normal states and identify system states that drastically deviate from them. Any network condition that significantly varies from usual signals an assault. The anomaly detection system detects novel network assaults. Designing an anomaly detection system is difficult. The attack state will go

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 8 Aug 2023, pp: 633-644 www.ijaem.net ISSN: 2395-5252

unnoticed if normal state patterns are similar to anomalous state patterns. This increases false alarms. Therefore, a normal condition must be designed to maximize the detection rate while minimizing false alarms. If the normal condition is too broad, detection will suffer. However, a narrow normal state increases false alarms. The hybrid detection technique combines the versatility and capability of an anomaly detection system with the precision and reliability of misuse detection.

Two important things must be done to make a hybrid detection system that works well and is accurate: (i) finding the best misuse or anomaly detection systems that can be combined with anomaly detection systems to make hybrid detection possible, and (ii) integrating the two systems in the best way to balance the rate of detection and the rate of false alarms while still being able to find new attacks.

The application determines the misuse and anomaly detection methods used in the hybrid detection system architecture. Anomaly detection and abuse detection integration have been divided into four categories using a combinational technique. The categories are anomaly-misuse sequence detection, misuse-anomaly sequence detection, parallel detection, and complex mixture identification (Figure 4). The complicated combination model is application-specific.

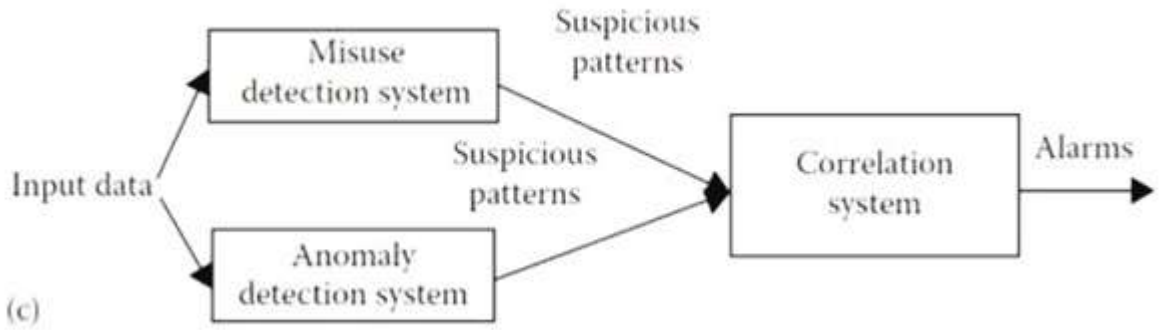

Figure 4. The hybrid parallel detection system.

VII. THE EVALUATION OF THE PROPOSED SYSTEM

We utilized NSL-KDD dataset in this work. The NSL-KDD dataset was enhanced to overcome KDD99's constraints. The dataset's website is public. The NSL-KDD data collection has 125,973 training examples and 22,544 test cases with 41 characteristics, 38 consistent, and 3 categorical. Two of the 23 training classes and 38 test classes are just for training, and 17 are odd for test knowledge. Classification is more complicated by class distribution variance. Training/testing classes are related to Normal, PROBE, R2L, U2R, or DoS. Each categorization except Normal is an incursion that assumes no abnormalities. IDS is still beneficial in these classes, which are very imbalanced and have enough cases to give more meaningful findings in each class.

Table.1. NSL-KDD Traffic Distribution

Traffic	Training	Test
Normal	67343	9711
DoS	45927	7458
Probe	11656	2754
R2L	995	2421
U2R	52	200
Total	125973	22544

The suggested model is tested using the aforementioned criteria based on dataset assaults. Accuracy is accurate identification, DR is the classifier's attack detection rate, FAR is the fraction of normal examples misclassified, and recall is the model's attack returns. Attacks are correct based on accuracy. Testing various result measures' performance must validate the HSOA model and compare it to alternative techniques.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 8 Aug 2023, pp: 633-644 www.ijaem.net ISSN: 2395-5252

Table.2. Proposed model result on traffic distribution of dataset

Attack	ACC	FAR	Precision	Recall
DoS	95.78	2.22	98.80	99.70
Probe	97.47	2.78	98.34	98.19
R2L	95.33	9.40	97.31	98.56
U2R	99.52	5.89	99.76	99.68

The NSL-KDD dataset's HSOA technique performance is shown in Table.2. Where each assault class is evaluated for accuracy, precision, recall, and FAR. Table.3 shows the effectiveness evaluation of the proposed model compared to various machine learning methods without and with HSOA optimization.

Table.3. Classification Accuracy Comparison

Method	Normal	DoS	R2L	Probe	U2R
SVM	95.89	81.44	83.48	87.77	78.45
SVM-HSOA	99.8	88.80	90.65	89.98	86.90
PCA	90.11	90.11	91.90	90.34	88.40
PCA-HSOA	96.09	96.78	94.98	95.14	92.10
Linear regression	92.00	90.22	91.67	90.09	80.40
Linear regression -HSOA	98.16	93.67	93.39	94.78	87.43
Logistic regression	89.04	81.56	80.78	82.65	80.56
Logistic regression-HSOA	94.76	90.35	89.67	90.54	88.33
DT	93.10	92.63	87.90	88.12	81.39
DT-HSOA	98.45	97.31	92.94	96.47	89.57
NB	88.23	86.50	84.87	88.23	76.80
NB-HSOA	93.09	90.22	92.95	94.20	85.90
KNN	94.06	90.45	91.90	90.60	89.06
KNN-HSOA	96.10	96.76	95.78	97.89	93.11
RF	97.54	90.09	90.65	90.45	80.55
RF-HSOA	99.32	93.95	95.80	94.54	87.39
DR	90.33	83.76	82.09	81.34	80.42
DR-HSOA	94.51	90.34	90.70	89.56	89.90
GB	90.00	90.54	91.45	90.09	84.12
GB-HSOA	92.22	94.65	95.62	98.80	89.50
Kmeans	93.89	90.12	90.32	92.98	91.89
Kmeans-HSOA	97.43	96.60	97.90	99.67	98.89

Where:DT is Decision tree, NB is Naive Bayes algorithm, RF is Random Forest, DR is Dimensionality Reduction, GB is Gradient Boosting.

The proposed HSOA model outperforms in most of the performance compared with the other techniques. For normal class, the proposed model obtained up to 10% more detection for all classes.

VIII. CONCLUSION

In this chapter, we covered machine learning and data mining methods for abuse and anomaly detection system design. A few well-known systems in the scientific community have been briefly reviewed. We also addressed the system's merits and downsides in relation to its uses and implementation in real-world networks.

Training data for classification-based algorithms must be balanced with normal and attack traffic information, therefore we were used the NSL-KDD for this purpose. It is ideal to have a wide range of attack traffic data, including innovative assaults, although it may not be practicable. Labeling data is required, with attack and regular traffic data clearly differentiated. Anomaly detection-based algorithms require training data labeling, like classification-based approaches used in abuse detection.

A proposed optimized anomaly detection technique assumes regular and abnormal traffic as training data. The detection accuracy was greatly

enhanced by training the method of detection entirely on traffic data with different ML algorithms and proposed hybrid swarm optimization. Result of proposed system deals to the enhancement in detection classifiers results.

Real-time detection is required for a real-world intrusion detection system in a high-speed, high-volume data environment. The proposed optimized ML techniques are scalable and require all training data to be in memory during training. This constrains model size. The scalability, performance, detection rate, and rate of false positives of anomaly detection systems are best as shown in the results above.

REFERENCES

[1]. Y. Lu and L. D. Xu, "Internet of things (iot) cybersecurity research: A review of current research topics," IEEE Internet of Things Journal, vol. 6, no. 2, pp. 2103–2115, Apr. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8462745

[2]. R. Roman, J. Zhou, and J. Lopez, "Internet of things (iot) cybersecurity research: A review of current research topics," Elsevier Computer Networks, vol. 57, no. 10, pp.2266–2279, Jul. 2018. [Online]. Available: https://doi.org/10.1016/j.comnet.2012.

[3]. 12.0183.Abraham A, Grosan C, Martin-Vide C. Evolutionary design of intrusion detection programs. International Journal of Network Security. 2007;4(3):328-339. DOI: 10.6633/IJNS.200705.4(3).12

[4]. Cannady J. Artificial neural networks for misuse detection. In: Proceedings of the National Information Systems Security Conference (NISSC'98). Washington, DC; 6-9 October 1998. pp. 441-454

[5]. Mukkamala S, Janoski G, Sung AH. Intrusion detection using neural networks and support vector machines. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN'02). Honolulu, HI; 12-17 May 2002. pp. 1702-1707. DOI: 10.1109/IJCNN.2002.1007774

[6]. Kruegel C, Toth T. Using detection trees to improve signature-based intrusion detection. In: Proceedings of the 6th International Workshop on Recent Advances in Intrusion Detection. Pittsburgh, PA; 8-10 September 2003. pp. 173-191. DOI: 10.1007/978-3-540-45248-5_10

[7]. Chebrolu S, Abraham A, Thomas JP. Feature deduction of intrusion detection systems. Computers & Security. 2005;24:295-307. DOI: 10.1016/j.cose.2004.09.008

[8]. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Machine Learning. 1992;9:309-347. DOI: 10.1007/BF00994110

[9]. Verma T, Pearl J. An algorithm for deciding if a set of observed independencies has a causal explanation. In: Proceedings of the 8th International Conference on Uncertainty in Artificial Intelligence. Stanford, CA; July 1992. pp. 323-330. DOI: 10.1016/B978-1-4832-8287-9.50049-9

[10]. Pearl J, Wermuth N. When can association graphs admit a causal interpretation? In: Proceedings of the 4th International Workshop on Artificial Intelligence and Statistics. Fort Lauderdale, FL; 1993. pp. 141-150. DOI: 10.1007/978-1-4612-2660-4_21

[11]. Schultz MG, Eskin E, Zadok E, Stolfo SJ. Data mining methods for detection of new malicious executables. In: Proceedings of IEEE Symposium on Security and Privacy (S&P'01). Oakland, CA. Anaheim, CA; 14-16 May 2000. DOI: 10.1109/SECPRI.2001.924286

[12]. Ghosh AK, Schwartzbard A, Schatz M. Learning program behavior profiles for intrusion detection. In: Proceedings of the 1st USENIX Workshop on Intrusion Detection and Network Monitoring. Santa Clara, CL; 9-12 April 1999. pp. 51-62

[13]. Gong F. Deciphering Detection Techniques: Part II. Anomaly-Based Intrusion Detection. Santa Clara, CA, USA: White paper, Mcafee Network Security Technologies Group; 2003

[14]. Eskin E, Arnold A, Prerau M, Portnoy L, Stolfo S. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In: Jajodia S, Barbara S, editors. Applications of Data Mining and Computer Security. Dordrecht: Kluwer; 2002. pp. 77-101. DOI: 10.7916/D8D50TQT

[15]. Lee W, Stolfo SJ. Data mining approaches for intrusion detection. In: Proceedings of the 7th USENIX Security Symposium. San Antonio, TX; 26-29 January 1998. DOI: 10.7916/D86D60P8

[16]. Apiletti D, Baralis E, Cerquitelli T, D'Elia V. Characterizing network traffic by means of the NetMine framework. Computer Networks. 2009;53(6):774-789. DOI: 10.1016/j.comnet.2008.12.011

[17]. Mannila H, Toivonen H. Discovering generalized episodes using minimal occurrences. In: Proceedings of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining. Portland, OR: P. ACM; August 1996. pp. 146-151

[18]. Luo J, Bridges SM. Mining fuzzy association rules and fuzzy frequency episodes for intrusion detection. International Journal of Intelligent Systems. 2000;15(8):687-703

[19]. tcpdump website. Available from: https://www.tcpdump.org

[20]. Ghosh AK, Wanken J, Charron F. Detecting anomalous and unknown intrusions against programs. In: Proceedings of the 14th Annual Computer Security Applications Conference (ACSAC'98). Phoenix, AZ; 7-1 December 1998. DOI: 10.1109/CSAC.1998.738646

[21]. L. Leenen and T. Meyer, "Artificial intelligence and big data analytics in support of cyber defense," Developments in Information Security and Cybernetic Wars, pp. 42–63, jan2019.

[22]. W. Ding, X. Jing, Z. Yana, and L. T. Yang, "Artificial intelligence and big data analytics in support of cyber defense," Information Fusion, vol. 51, pp. 129–244, dec2018. [Online]. Available: https://doi.org/10.1016/j.inffus.2018.12.001

[23]. Hu WJ, Liao YH, Vemuri VR. Robust support vector machines for anomaly detection in computer security. In: Proceedings of the International Conference on Machine Learning (ICMLA'03). Los Angeles, CL: CSREA; 23-24 June 2003. pp. 161-167

[24]. Liao YH, Vemuri VR. Use of k-nearest neighbor classifier for intrusion detection. Computers & Security. 2002;21(5):439-448. DOI: 10.1016/S0167-4048(02)00514-X

[25]. Warrender C, Forrest S, Pearlmutter B. Detecting intrusions using system calls: Alternative data models. In: Proceedings of IEEE Symposium on Security and Privacy. Oakland, CA: IEEE; 10-14 May 1999. pp. 133-145. DOI: 10.1109/SECPRI.1999.766910

[26]. Qiao Y, Xin XW, Bin Y, Ge S. Anomaly intrusion detection method based on HMM. Electronics Letters. 2002;38(13):663-664. DOI: 10.1049/el:20020467

[27]. Wang W, Guan X, Zhang X, Yang L. Profiling program behavior for anomaly intrusion detection based on the transition and frequency property of computer audit data. Computers & Security. 2006;25(7):539-550. DOI: 10.1016/j.cose.2006.05.005

[28]. Sammut C, Webb GI, editors. Encyclopedia of Machine Learning. Boston, MA: Springer; 2011. DOI: 10.1007/978-0-387-30164-8

[29]. Li SA, Jain A, editors. Encyclopedia of Biometrics. Boston, MA: Springer; 2009. DOI: 10.1007/978-0-387-73003-5_592

[30]. Soule K, Salamatian K, Taft N. Combining filtering and statistical methods for anomaly detection. In: Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement. Berkeley, CA: ACM; 19-21 October 2005. pp. 331-344. DOI: 10.1145/1330107.1330147

[31]. D. M. Hawkins, "Identification of outliers," Monographs on statistics and applied probability, pp. 1–194, 1980. [Online]. Available: https://link.springer.com/content/pdf/10.1007/978-94-015-3994-4.pdf.

[32]. Portnoy L, Eskin E, Stolfo S. Intrusion detection with unlabeled data using clustering. In: Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA). Philadelphia, PA: ACM; November 2001. pp. 5-8. DOI: 10.7916/D8MP5904

[33]. Zhang J, Zulkernine M. Anomaly-based network intrusion detection with unsupervised outlier detection. In: IEEE International Conference on Communications. Istanbul, Turkey: IEEE; 11-15 June 2006. pp. 2388-2393. DOI: 10.1109/ICC.2006.255127

[34]. Zhang J, Zulkernine M, Haque A. Random forest-based network intrusion detection systems. IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews.

2008;38(5):649-659. DOI: 10.1109/TSMCC.2008.923876

[35]. Eskin E. Anomaly detection over noisy data using learned probability distribution. In: Proceedings of the 17th International Conference on Machine Learning (ICML'00). Stanford, CA: ACM; 29 June-2 July 2000. pp. 255-262. DOI: 10.7916/D8C53SKF

[36]. Ye N, Li X, Chen Q , Emran SM, Xu M. Probabilistic techniques for intrusion detection based on computer audit data. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans. 2001;31(4):266-274. DOI: 10.1109/3468.935043.

[37]. U. S. Shanthamallu, A. Spanias, C. Tepedelenlioglu, and M. Stanley, "A brief survey of machine learning methods and their sensor and iot applications," 2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA), pp. 1–8, Aug. 2017.

[38]. P. Suresh, J. V. Daniel, V. Parthasarathy, and R. H. Aswathy, "A state of the art review on the internet of things (iot) history, technology and fields of deployment," in 2014 International Conference on Science Engineering and Management Research (ICSEMR), 11 2014, pp. 1–8.

[39]. Yamanishi K, Takeuchi J, Williams G, Milne P. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. Data Mining and Knowledge Discovery. 2004;8(3):275-300. DOI: 10.1023/B:DAMI.0000023676.72185.7c

[40]. Mahoney MV, Chan PK. Learning nonstationary models of normal network traffic for detecting novel attacks. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada: ACM; 23-26 July 2002. pp. 376-386. DOI: 10.1145/775047.775102

[41]. Ye N, Emran SM, Chen Q , Vibert S. Multivariate statistical analysis of audit trails for host-based intrusion detection. IEEE Transactions on Computers. 2002;51(7):810-820. DOI: 10.1109/TC.2002.1017701

[42]. Lakhina A, Crovella M, Diot C. Mining anomalies using traffic features distributions. Computer Communication Review. 2005;35(4):217-228. DOI: 10.1145/1090191.1080118

[43]. Lakhina A, Crovella M, Diot C. Diagnosing network-wide traffic anomalies. In: Proceedings of the 2004 International Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM'04). 2004. pp. 219-230. DOI: 10.1145/1015467.1015492

[44]. Ringberg H, Soule A, Rexford J, Diot C. Sensitivity of PCA for traffic anomaly detection. Performance Evaluation Review. 2007;35(1):109-120. DOI: 10.1145/1269899.1254895

[45]. Lee W, Xiang D. Information-theoretic measures for anomaly detection. In: Proceedings of 2001 IEEE Symposium on Security and Privacy. Oakland, CA; 14-16 May 2000. DOI: 10.1109/SECPRI.2001.924294

[46]. Zhang J, Zulkernine M. Anomaly based network intrusion detection with unsupervised outlier detection. In: Proceedings of the IEEE International Conference on Communications (ICC'06). Istanbul, Turkey; 11-15 June 2006. DOI: 10.1109/ICC.2006.255127

[47]. Zhang J, Zulkernine M, Haque A. Random-forest-based network intrusion detection systems. IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews. 2008;38(5):649-659. DOI: 10.1109/TSMCC.2008.923876

[48]. Barbara D, Couto J, Jajodia S, Wu N. ADAM: A testbed for exploring the use of data mining in intrusion detection. In: Proceedings of the ACM SIGMOD. Santa Barbara, CL; May 2001. DOI: 10.1145/604264.604268

[49]. Zhang J, Zulkernine M. A hybrid network intrusion detection technique using random forests. In: Proceedings of the 1st International Conference on Availability, Reliability, and Security (ARES'06). Vienna, Austria: IEEE; 20-22 April 2006. DOI: 10.1109/ARES.2006.7

[50]. Anderson D, Frivold T, Valdes A. Next-generation intrusion detection expert system (NIDES) – A summary. Technical Report SRI-CSL-95-07, SRI; 1995

[51]. Agrawal R, Gehrke J, Gunopulos D, Raghavan P. Automatic subspace clustering of high dimensional data for data mining applications. In: Proceedings

of ACM SIGMOD. Seattle, WA: ACM; 1998. pp. 94-105. DOI: 10.1145/276305.276314

[52]. Sen J, Sengupta I. Autonomous agent-based distributed fault-tolerant intrusion detection system. In: Proceedings of the 2nd International Conference on Distributed Computing and Internet Technology (ICDCIT'05). Vol. 3186. Bhubaneswar, India: Springer, LNCS; 22-24 December 2005. pp. 125-131. DOI: 10.1007/11604655_16