An Innovative Deep Learning Method for Identifying Anomalies and Preventing Intrusions in Networked Systems

Fnu Ziauddin*

Computer Network Architect, Dallas, Texas, USA

Research Article

Received: 30-Jan-2023, Manuscript No. GRCS-23-123147; **Editor assigned:** 02-Feb-2023, PreQC No. GRCS-23-123147(PQ); **Reviewed:** 16-Feb-2023, QC No. GRCS-23-123147; **Revised:** 23-Feb-2023, Manuscript No. GRCS-23-123147(R); **Published:** 02-Mar-2023, DOI: 10.4172/ 2229-371X.14.1.011.

***For Correspondence:** Fnu Ziauddin, Computer Network Architect, Dallas, Texas, USA

E-mail: uziad451@gmail.com

Citation: Ziauddin F. An Innovative Deep Learning Method for Identifying Anomalies and Preventing Intrusions in Networked Systems. J Glob Res Comput Sci. 2023;14:011.

ABSTRACT

Over the last two to three decades, cyber security has grown significantly in importance due to the remarkable progress and use of computer networks. Large amounts of data are transferred and received across networks, and as these networks have grown, so have the scope and sophistication of assaults. As a result, data is vulnerable to assault while it is transported and stored. To guarantee network security and prevent malware assaults, a strong networked intrusion detection system (IDS) is necessary. In contrast, an IDS is seen as essential to breaches in the availability, privacy, integrity, and confidentiality of data and other resources within the context of network security frameworks. An intrutraffic andon system watches what happens on the network, examines the network traffic, and notifies the system if it detects any odd activity or incursion. Protecting an attacker's network requires anomaly detection as a critical component. Finding threats inside a network by examining its behavior pattern was crucial for many researchers and application frameworks in both IPv4 and IPv6 networks. An effective data mining approach, like machine learning, must be employed to find anomalies. The procedure of gathering data is becoming more significant in this study as it relates to the investigation of the anomalies. For testing purposes, we have used the Knowledge Discovery and Data Mining (KDD) Cup network traffic dataset, which will imitate real-time attack behavior. Panda's data frame has the CSV file loaded, displaying the data in a tabular format. 41 features totaling 10,000 occurrences were used for 4 distinct classes, including "dos", "normal," "probe," and "r2l." For our investigation in this study project, we used two machine learning (ML) approaches and one deep learning methodology. The "Fastai Library" from deep learning has been used by us for intrusion detection categorization. Nonetheless, we have used the Random Forest (RF) and Decision Tree (DT) techniques for analysis in machine learning. Based on accuracy, we have contrasted the deep learning and machine learning models. The Fastai Library has a 92% accuracy rate, Decision Tree has an 82% accuracy rate, and Random Forest has an 84% accuracy rate. Thus, our analysis of accuracy indicates that the deep learning (DL) method may improve the performance of the intrusion detection system (IDS).

Keywords: Anomaly detection; Machine Learning; IPv4 & IPv6; Network traffic dataset; Real-time attack

INTRODUCTION

The internet and enterprise networks are crucial for creating and supporting new business opportunities. However, this makes them more prone to intrusions and attacks. The variety and ever-changing nature of these attacks demand a flexible approach to network security. A flexible defense system is necessary, one that can analyze large volumes of network traffic and accurately detect various attacks. Anomaly-based intrusion detection is a useful method for identifying both known and new attacks in intrusion detection systems [1]. It works by continuously monitoring and modeling normal network behavior, then identifying potential threats through deviations from this norm. Anomalies are critical in this context because they signal rare but serious events. For instance, unusual traffic patterns might indicate an attack and unauthorized data transmission [2]. Anomalies can be categorized into three types: point anomalies, contextual anomalies, and collective anomalies, each correlating with different types of network attacks like DoS, Probe, U2R, and R2L. Differentiating these anomalies is essential for identifying and classifying network attack types [3]. Anomaly-based NIDS (Network Intrusion Detection Systems) must adapt to dynamic network environments, recognizing new protocols and behaviors. In anomaly-based NIDS, the system learns from 'normal' network traffic to create a model. This model is then used to classify new activities as normal or anomalous. Effective models should adapt to and cope with dynamic network environments, essentially requiring a self-learning system [4]. Deep learning, a subset of machine learning, uses multiple layers of information processing for unsupervised learning and pattern classification. While deep learning has been successful in areas like computer vision and natural language processing, its potential in intrusion detection is still underutilized [5]. This research study work focuses on leveraging deep learning's strengths in self-learning and big data analysis for anomaly-based IDS.

A major challenge in this field is the lack of representative datasets. We plan to use the KDD Cup network traffic dataset, which categorizes cyber-attacks into nine types based on their behavior, as shown in Table 1. This addresses the dataset shortage highlighted by [6], a key obstacle for such detection systems. The KDD Cup network traffic dataset categorizes cyber-attacks into several types based on their behaviors. Generic attacks disrupt systems by causing hash function collisions. Exploits manipulate unexpected behavior in networks by leveraging vulnerabilities [7]. DOS attacks block legitimate access to services by overwhelming networks with excessive traffic. Reconnaissance attacks involve adversaries gathering information to find system weaknesses. Analysis attacks are where intruders exploit technical vulnerabilities after gaining network access [8].

Table 1. Comprehensive classification of cyber-attack types: Understanding threats in network security through the KDD cup dataset.

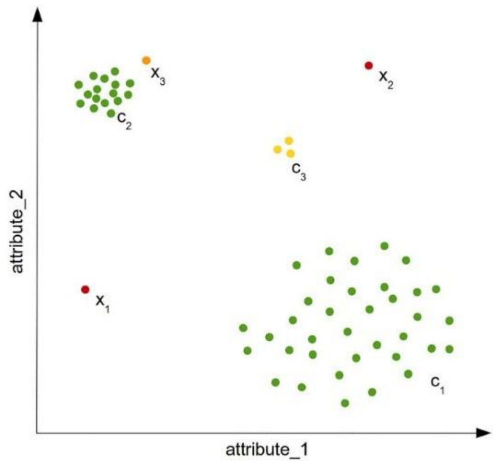
No	Type	Description
1	Generic	An attack that causes a collision by using the hash function against every blocked cipher
2	Exploit	A sequence of commands that cause unexpected behavior on the network by exploiting a vulnerability or bug
3	Fuzzers	An attack where an adversary tries to find security loopholes or bugs in network, operating system, or software by randomly feeding data
4	DOS	An attack that prevents a legitimate user from accessing services by constantly forwarding packets on the network or a specified host
5	Reconnaissance	A cyber-attack where an adversary tries to find information about a targeted system or network to uncover weaknesses or vulnerabilities
6	Analysis	Focuses on intrusion of computer and network where an adversary gains access to a system and exploits vulnerabilities using technical abilities
7	Backdoor	A covert method by which an attacker bypasses normal authentication in a network
8	Shellcode	An attack targeting a vulnerable process running on another machine on a local network using shellcode as a payload to exploit the compromised machine
9	Worms	A detached intrusion that replicates itself and spreads through network connections and downloading

Backdoor attacks bypass standard authentication, shellcode attacks target vulnerable processes for exploitation, and worms self-replicate to spread through networks [9]. Each type of attack requires specific attention to ensure robust network security. Approaches that are being used widely for intrusion detection also includes Machine learning approaches which are the study of algorithms that learnt efficiently through training data for the identification of Cyber-attacks or any authorized access at both host and network level. These algorithms give assistance in extracting useful knowledge from huge amounts of data.

Anomaly detection

Security systems known as intrusion detection systems can gather and examine data from a variety of system and network sources to identify any behaviour that would indicate an attack or unauthorized access to the system. The term "intrusion detection" describes the process of identifying malevolent activity in a host or network, such as hacking, penetrations, and other computer misuse [10]. Since an intrusion deviate from the system's typical behaviour, anomaly detection methods may be used in the intrusion detection field. An IDS may be host- or network-based, depending on the information source that is taken into consideration. Events relating to OS information, such as system calls and process IDs, are analyzed by a host-based intrusion detection system [11]. The abnormal subsequences of the traces represent the intrusions. The aberrant subsequences result in harmful programmes, unlawful activity, and policy violations. The sequential structure of data necessitates the use of anomaly detection methods for host-based intrusion detection [12]. The methods need to calculate sequence similarity or model the sequence data. Using system call traces, anomalies may be found using two popular methods: frequency-based and short sequence-based techniques. Regression and clustering are among the most often used methods for distinguishing abnormalities, as Figure 1 illustrates.

Figure 1. Anomaly in 2-Dimensional dataset.



The image is a scatter plot with two axes labeled "attribute_1" and "attribute_2," which represent two variables or features of a dataset. The plot shows a large cluster of green data points that likely represent normal instances within the dataset. There are also three distinct colored data points labeled X1, X2, and X3, which are separated from the main cluster, indicating that they could be anomalies or outliers. Additionally, there are three labeled points C1, C2, and C3, possibly representing the centroids of clusters if a clustering algorithm like K-means has been used. In the context of identifying anomalies and preventing intrusions in networked systems, the plot could illustrate how a deep learning model distinguishes between normal operations and potential security threats, with the separated points potentially representing unusual activity that could signify an intrusion. A network-based IDS looks at network events like how much traffic there is, IP addresses, the services being used, and the protocols in play. It watches the network's usual activity and flags anything that doesn't match as a potential intrusion 13. This kind of data is complex because it has many different types of information, both categorical (like service types) and numerical (like traffic volume). The system must be fast and efficient to process all this data because it often comes in real time. We usually know what normal data looks like, but we don't always have clear examples of intrusions. This is why methods that don't need as many labeled examples of intrusions are often used. In network-based IDS, common ways to spot anomalies include methods based on classification, statistics, information theory, and grouping data into clusters.

MATERIALS AND METHODS

The research study presents a systematic technique for anomaly identification in IPv4/IPv6 networks, starting with network traffic data collecting. The basis for behaviour profiling is this data, which is used to analyze common network behaviors and patterns to create a baseline of what is deemed normal. The pre-treatment of the data, which includes cleaning, normalizing, and dimensionality reduction to guarantee that the data is in a good form for analysis, is the next essential step. The complex process of feature selection, which involves selecting important characteristics suggestive of network behaviour, comes next. These characteristics are essential since they will be used to identify departures from the norm that might imply the existence of anomalies. The next step is to design control variables to fine-tune the detection method with the goal of effectively identifying threats vs benign abnormalities. The core of the approach is the anomaly detection procedure, in which the system examines network data in real-time to identify abnormalities by using pre-existing behaviour profiles and certain attributes. The system's capacity to adjust and learn over time, using machine learning algorithms to change in response to shifting network traffic patterns, improves this ongoing monitoring. The capacity to adapt dynamically is crucial for preserving the anomaly detection system's accuracy and relevance in the face of the constantly evolving network security threat environment.

A hybrid intrusion detection model was created for the research project by combining several machine learning methods. The model uses Minimal Sequence Optimization (MSO) for improved operation sequence, an optimised K-means clustering technique for effective data training, and Extreme Learning Machine (ELM) for quick learning. The Deep Learning models showed greater accuracy, especially a 92% accuracy rate using the Fastai Library compared to 81.0303% for DT and 83.1515% for RF. Decision Trees (DT) and Random Forest (RF) approaches were also utilized.

Figure 2. Methodology diagram of intrusion detection in IPv4/IPv6 networks.

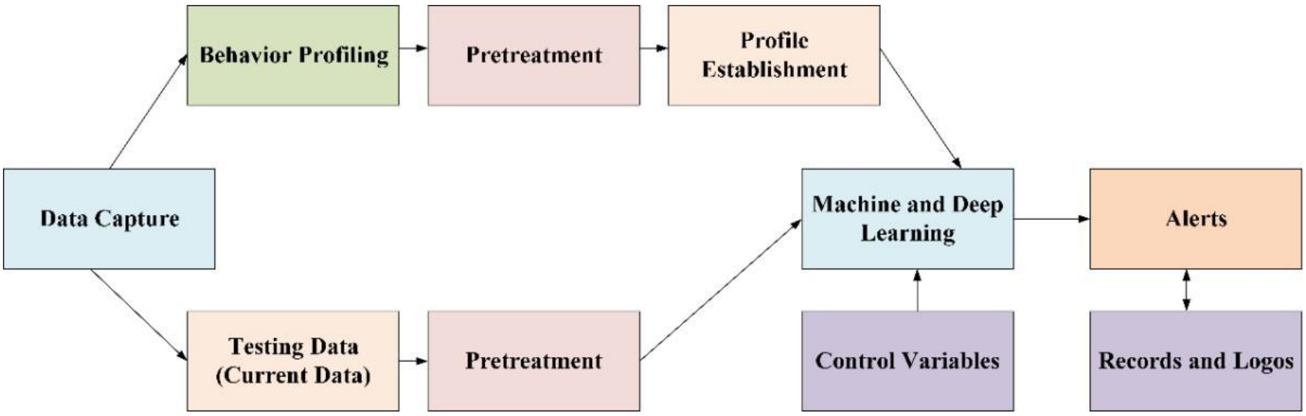
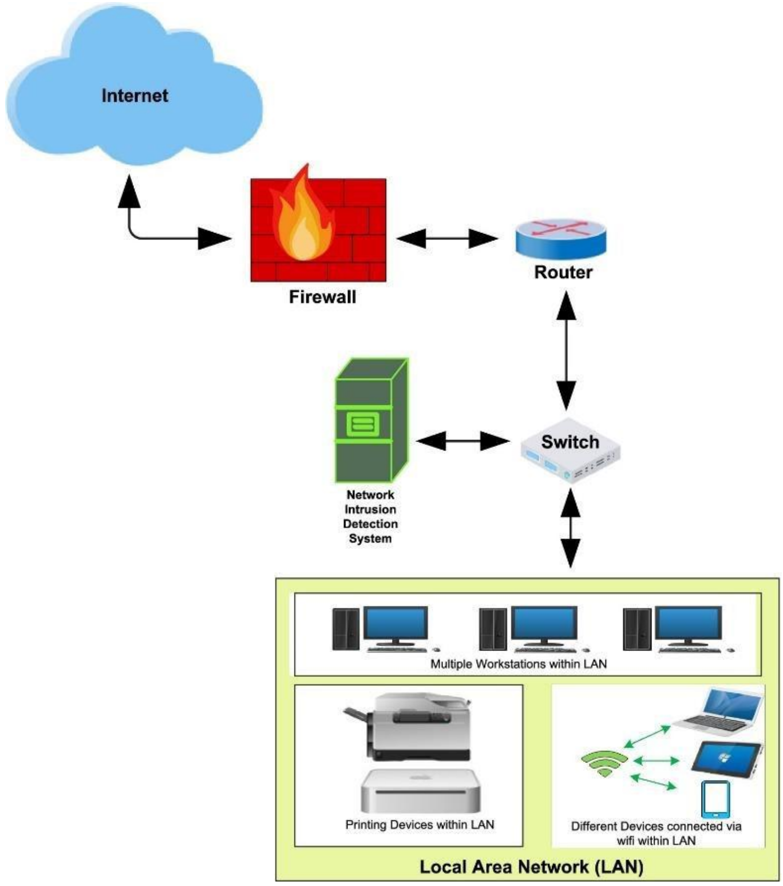


Figure 2 outlines the process flow of an anomaly detection system in network security. The process begins with "Data Capture," where network traffic data is collected. This data is then subjected to "Behavior Profiling" to understand normal network activity. Concurrently, "Testing Data" representing current network activity is also captured. Both sets of data undergo "Pretreatment" to clean and standardize them, making the data suitable for analysis. After pretreatment, a "Profile Establishment" occurs for the behavior-profiled data, creating a baseline of normal activity. This baseline is then used alongside the control variables within the "Machine and Deep Learning" framework to analyze the pretreated testing data. If anomalies are detected, the system generates "Alerts." Furthermore, all the events and outcomes are logged in "Records and Logos," creating a history of detections and system actions. This systematic approach integrates continuous learning and adjusting, as the machine and deep learning algorithms refine the normal behavior profile over time based on the data and feedback received.

Classification of IDS based on ML

An IDS, short for Intrusion Detection System, combines the ideas of 'intrusion' and a 'detection system.' 'Intrusion' means any unauthorized entry into a computer or network, which could harm its safety, privacy, or functioning. A 'detection system' is a tool that spots such illegal activities. Therefore, an IDS is a security device that keeps an eye on computers and network traffic. It looks for any strange actions that break the security rules and could affect privacy, trustworthiness, or working. If it finds harmful behavior, the IDS alerts the computer or network managers. Figure 3 shows how a Network Intrusion Detection System (NIDS) is set up in a passive way. It connects to a network switch using port mirroring, which copies all network traffic to the NIDS for monitoring and finding intrusions. A NIDS can also be placed between the firewall and network switch, checking all traffic that goes through the IDS.

Figure 3. Passive deployment of network-based intrusion detection system [13].



RESULTS AND DISCUSSION

In the "Results" section, the focus is on the critical need for high-level security to facilitate stable and reliable communication between various entities. The study acknowledges the inherent risks of network connectivity, including the potential for abuses like intrusions, which are common threats to computer systems worldwide. Specifically, the paper addresses the detection and prevention of cyber-attacks such as Denial of Service (DoS), User Remote attacks (R2L), and Probing. The research incorporates the "Fastai Library" for deep learning to enhance intrusion detection capabilities. Additionally, Decision Tree (DT) and Random Forest (RF) techniques from machine learning are utilized. The KDD Cup 1999 network traffic dataset, which is designed to mimic real-time attack behavior, serves as the basis for testing. The dataset is loaded into a panda Data-Frame and consists of 41 attributes with 10,000 instances spread over four distinct classes: ‘dos’, ‘normal’, ‘probe’, and ‘r2l’. A comparative analysis of machine learning and deep learning models is conducted to evaluate their accuracy in anomaly detection within IPv4/IPv6 Networks. The experimentation is performed using Python, with Google Co-Lab selected as the IDE for this research. The analysis demonstrates the potential of these computational techniques in recognizing and responding to security threats in networked environments.

Figure 4. The confusion matrix for fastai library based on the testing set of data.

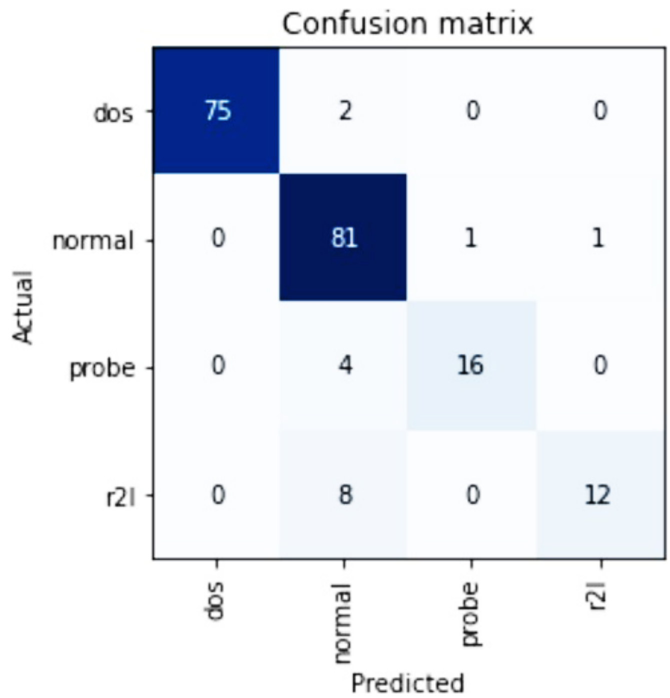
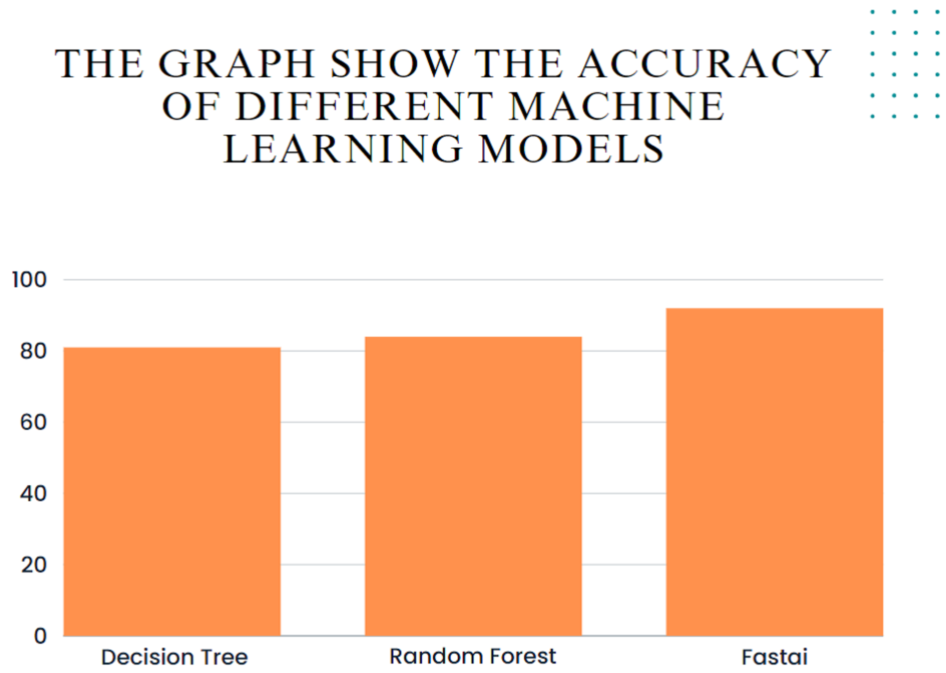


Illustration 4 presents the confusion matrix obtained during the testing phase using the Fastai library. The dataset was partitioned such that 98% was utilized for training purposes, while the remaining 2% was served for testing. The matrix's diagonal entries depict the instances that were accurately classified, such as dos-to-dos classifications. Non-diagonal entries represent the instances of misclassification, for example, where a dos instance was incorrectly classified as normal, among other misclassifications. This matrix serves as a visual representation of the model's performance, highlighting its precision in correctly identifying and categorizing different types of network traffic under test conditions. The provided confusion matrix is a visual representation of the performance of a classification model, detailing the accuracy of predictions against actual labels. The matrix's diagonal entries reveal a high number of correct predictions: 'dos' with 75, 'normal' with 81, 'probe' with 16, and 'r2l' with 12, indicating the model's effectiveness at identifying true positives for each class. The non-diagonal entries show misclassifications, such as 'dos' incorrectly classified as 'normal' twice, and 'normal' mislabeled as 'probe' and 'r2l' once each. Notably, 'probe' is also misclassified as 'normal' four times, and 'r2l' is misclassified as 'normal' eight times, suggesting areas where the model could be improved. There are no misclassifications from 'probe' or 'r2l' into 'dos', indicating certain consistencies in the model's predictive behavior. This confusion matrix effectively summarizes the model's classification strengths and weaknesses, providing insights for further refinement.

Table 2. The comparison table for the machine learning vs deep learning.

No	Algorithm's	Accuracy
1	Decision Tree	82%
2	Random Forest	84%
3	Fastai	92%

Figure 5. The bar graph show the accuracy of different machine learning models.



The bar chart in figure 5 displays the comparative accuracy of three different machine learning models: Decision Tree, Random Forest, and Fastai. It presents a visual comparison where each model's accuracy is represented by the height of the bar, with the scale on the vertical axis ranging from 0 to 100 percent. The Decision Tree and Fastai models exhibit similarly high bars, suggesting comparable levels of accuracy, while the Random Forest model shows a slightly lower bar, indicating a marginal reduction in accuracy. The precise numerical values of the accuracy percentages are not discernible from the image. Such a graphical representation is a common method for succinctly illustrating the performance metrics of various predictive models in machine learning.

CONCLUSION

Anomaly detection is essential for protecting network systems against cyber threats. Identifying anomalies by analyzing behavioral patterns in network traffic is a key focus area for researchers and is crucial for both IPv4 and IPv6 networks. To detect anomalies effectively, the application of advanced data mining techniques, such as machine learning algorithms, is vital. In this study, the data collection phase is always crucial for analyzing anomalies. The KDD Cup 1999 network traffic dataset, known for its ability to emulate real-time cyber-attacks, was utilized for testing purposes. The data was organized into a pandas Data-Frame, displaying 41 attributes across 10,000 instances categorized into four classes: 'dos', 'normal', 'probe', and 'r2l'. The research employed one deep learning technique, using the Fastai Library for intrusion classification, and two machine learning techniques: Decision Tree (DT) and Random Forest (RF) for analytical comparison. The comparative analysis focused on the accuracy of these models, with the Fastai Library achieving a 92% accuracy rate, the Decision Tree reaching 81.0303%, and the Random Forest scoring 83.1515%. These results indicate that the Intrusion Detection System (IDS) performs more effectively when utilizing deep learning algorithms.

REFERENCES

1. Mukkamala S, et al. Cyber security challenges: Designing efficient intrusion detection systems and antivirus tools. Enhancing Computer Security with Smart Technology. 2005:125-163.

2. Buczak AL, et al. A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Commun Surv Tutor. 2016;18:1153-1176.

3. Haq NF, et al. Application of machine learning approaches in intrusion detection system: A survey. Int J Adv Res Artif Intell. 2015;4:1-9.

4. Bhuyan MH, et al. Towards generating real-life datasets for network intrusion detection. Int J Netw Secur. 2015;17:683-701.

5. Ring M, et al. A survey of network-based intrusion detection data sets. Comput Secur. 2019;86:147-167.

6. Sommer R, et al. Outside the closed world: On using machine learning for network intrusion detection. 2010 IEEE Symposium on Security and Privacy. 2010:305-316.

7. Desale KS, et al. Genetic algorithm based feature selection approach for effective intrusion detection system. 2015 International Conference on Computer Communication and Informatics. 2015:1-6.

8. Agrawal S, et al. Survey on anomaly detection using data mining techniques. Procedia Comput Sci. 2015;60:708-713.

9. Medel JR, et al. Anomaly detection in video using predictive convolutional long short-term memory networks. Arxiv. 2016:1-18.

10. Anwar S, et al. Response option for attacks detected by intrusion detection system. 2015 4th International Conference on Software Engineering and Computer Systems. 2015:195-200.

11. Elejla OE, et al. Intrusion detection systems of ICMPv6-based DDoS attacks. Neural Comput Appl. 2018;30:45-56.

12. Malhotra P, et al. Long short-term memory networks for anomaly detection in time series. In Esann. 2015:89.

13. Ahmad Z, et al. Network intrusion detection system: A systematic study of machine learning and deep learning approaches. Trans Emerg Telecommun Technol. 2021;32:e4150.