International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 11 Issue 8, August-2022, Impact Factor: 7.957

An AI Enabled an Intrusion Detection System for Proactive Cloud Infrastructure Security

Suresh S[1], Dr. Manisha[2]

[1]Research Scholar, Department of Computer Science & Engineering, JS University, Shikohabad, UP
[2]Assistant Professor Supervisor, Department of Computer Science & Engineering, JS University, Shikohabad, UP

ABSTRACT

In this day and age, the Intrusion Detection System (IDS) is one of the most well-known and widely employed systems that is now available. The preservation of sensitive data, the provision of an additional layer of security, and the prevention of illegal access to networks are the key purposes of this application. Data that is sent over a network is analyzed by intrusion detection systems (IDS), which then flag anything that might potentially be hazardous in order to ensure the safety of hosts and networks. In addition, alarm systems are designed to identify any action that is not typical. Numerous new sectors have evolved as a direct consequence of the fast spread of the Internet. Some examples of these new industries include big data, cloud computing, and the Internet of Things (IoTs). By increasing the amount of data that is created and sent throughout the network, this may be one of the factors that is contributing to the increase in the frequency of attacks.

In light of this, a great number of researchers have focused their attention on intrusion detection systems (IDS) and contributed to the effectiveness of these systems in preventing attacks and other threats associated with comparable threats. However, it is still likely that a significant amount of the data that is contained in network records is comprised of features that are irrelevant to the identification or classification of attacks. Therefore, professionals continue to have a difficult time deriving meaningful insights from this sort of network data and assessing whether or not the chosen qualities could increase the performance of immune system detection systems (IDS). In addition, breach detection systems need a substantial collection in order to effectively deal with the diverse range of threats. In order to enhance the speed and accuracy of the intrusion detection system (IDS), it is necessary to undertake the difficult process of determining the principal characteristics.

For the purpose of identifying threats and learning from previous data based on patterns, the current intrusion detection system (IDS) makes use of a wide range of machine learning, deep learning, and evolutionary algorithms. It is possible that the implementation of these strategies will be costly owing to the fact that they take into account all aspects of traffic at the same time. Yet, these strategies are successful. It is thus crucial to consistently address the problem of cutting expenditures without losing efficiency and the inclusion of features that are not necessary. However, in order to address these issues, the NSL-KDD and CICIDS2017 files were first analyzed using FSA. This was done in order to get rid of characteristics that were not required and to concentrate on those that were essential. For the purpose of satisfying this demand, it was necessary to design more efficient intrusion detection systems (IDS) that were capable of operating in very large networks at much lower prices. We examine and evaluate a variety of models that make use of FSA by making use of NSL-KDD datasets in order to enhance the detection engine of the intrusion detection system (IDS).

Keywords: Intrusion Detection System (IDS) Feature Selection Algorithm (FSA), Machine Learning, Network Security

INTRODUCTION

Over the course of the last several decades, the use of online services has garnered an excessive amount of popularity in the current world. These services are currently used by the majority of customers around the clock and from any place using a variety of electronic devices such as mobile phones, computers, tablets, and other similar devices. As a consequence of this, information that is either sensitive or important may be transmitted via these networks.

Another consequence of the ever-improving internet is the continuous flow of sensitive information between devices and data centers for the purposes of storage and retrieval.

As a result of these repercussions, the perpetrators of the attack have a window of time to deliver a number of strikes that might put the person or group that they are intended to target in jeopardy. An attacker might potentially exploit system security flaws using a number of different approaches that are considered to be state-of-the-art. Users who are not allowed to use the system may get access to it, which might result in the compromise of sensitive information or the compromising of their accounts. When it comes to protecting system administrators and security workers from the risks that are now present, modern security solutions are very necessary. Both the Internet of Things and big data are examples of new technologies that are adding to the ever-increasing flow of data.

As a consequence of this, the network is getting more crowded with data, which makes it more difficult, time-consuming, and slow to alter the attack profile. In addition, it is essential for data scientists, businesses, and marketers to have the capacity to filter through such enormous data sets in order to find information that is either relevant or useful. The volume of data created by these connections is becoming an increasingly relevant worry for scientists and academics who are concerned with network security. This is because the number of people using the internet continues to climb. In the field of network security, the study of stopping unauthorized individuals from getting access to computer systems or networks via the identification and repair of security flaws is referred to as network security. Firewalls and antivirus software are only two examples of the many different defensive solutions that have been created over the last twenty years to protect against threats such as denial-of-service (DoS), user-to-root (U2R), remote-to-local (R2L), probing, and others.

For this reason, it is very necessary to include core security measures in order to identify new types of assaults as well as harmful data or traffic that might potentially damage the system or network. This kind of instrument is known as an intrusion detection system, or IDS [1]. It is usual practice to refer to them as "IDSs." A mix of hardware and software technologies is used by an intrusion detection system (IDS) in order to collect, analyze, and identify data that is being received. The employment of these instruments allows for the identification of a variety of threats, including those that are fraudulent attacks, possible risks, and network and individual system obnoxiousness [2]. When sensitive data is sent across a network, it is the responsibility of an intrusion detection system, often known as an IDS, to safeguard it. Examining the intrusion detection system (IDS), interpreting its data via the use of mathematical or statistical techniques, and ensuring that it notifies network administrators and managers of any odd behavior are all necessary steps that must be taken in order to complete these duties and achieve these goals [3].

A. PROBLEM STATEMENT

Over the course of the last two decades, having access to internet-based services has grown much more important, especially after the introduction of COVID-19. It is possible that the rapid rise to notoriety may be attributed to the emergence of much improved Internet technologies. Electronic gadgets, such as laptops, tablets, smartphones, and other similar devices, are often used by individuals in order to gain access to these services in a timely manner and from any place. As a consequence of this, sensitive information is increasingly being sent over these networks as it is being transferred between different computers and data storage facilities.

As a result of this, criminals may conduct huge attacks that might put the business or its clients in jeopardy by way of bypassing security precautions. Attackers use a wide range of complex strategies in order to exploit weaknesses in computer systems. A wide variety of these strategies are described here. It is possible that sensitive data might be exploited, user accounts could be stolen, or unauthorised access could be gained to the system as a result of this. Experts and scientists are now concentrating on securing highly sensitive data and strengthening networks to mitigate the impact of these assaults. A solution has been found in the widespread use of intrusion detection systems, sometimes known as IDS. Entering data is analyzed by intrusion detection systems in order to evaluate whether or not it refers to activity on the system or across the network. The proliferation of tools like the internet, social media, and the Internet of Things has resulted in an increase in the amount of data that is being generated and transferred within the network.

This is a consequence of the rise in popularity of these tools. There is a possibility that network traffic may have a variety of effects, some of which could be annoying while others would be negligible. In order to address this problem, intrusion detection systems (IDS) that are effective will have methods for adding or deleting features, as well as a range of monitoring strategies. The importance of this cannot be overstated since it will prevent the system's processing power and working time from expanding. The development of speedy decision-making engines and models for the reduction or deletion of features was something that was done in order to solve this problem [17]. Using a single classifier or estimate is probably not the best way to compare and evaluate different models.

2. BACKGROUND

It is becoming more difficult to maintain the security of computer systems as a result of the growth of different kinds of networks (such as software-defined and wireless sensor networks) and the ever-changing nature of attack strategies. When these more recent networks were being set up, security elements were not first considered to be of the utmost importance. The majority of the time, the typical security mechanisms that are in place do not provide enough protection for these networks. Since this is the case, it is necessary to have an instant security system that is able to identify attempts to infiltrate a computer system. As a result of this need, the Intrusion Detection System (IDS) was developed, and it is now widely acknowledged as an essential component of security systems. Intrusion detection systems (IDS) are designed to search for and keep track of possible security breaches [19]. This is done in order to guarantee that the three fundamental principles of computer security—authentication, integrity, and confidentiality—are still in place. This is done with the intention of identifying any possible violations or hazards that may exist.

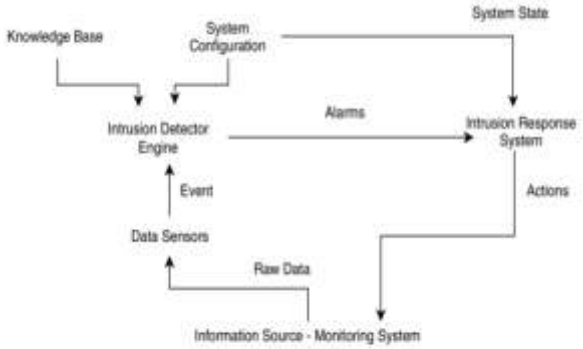

Figure 2.1 Intrusion detection and response technique [21] [22].

A. PARTS AND FUNCTIONS OF CUTTING-EDGE IDS

An Advanced Intrusion Detection System (IDS) is made up of either hardware or software, and its purpose is to monitor the network for any unusual behavior in order to detect possible threats. An Advanced Intrusion Detection System's key duties are to monitor networks, notify administrators of any suspicious behavior, and detect intrusions [26, 27]. These operations are intended to be performed simultaneously.

In the event that there is proof of misconduct When a monitoring system identifies potentially malicious behavior, for instance, it may decide to block traffic coming from the IP address that is being watched [26, 28]. According to what is said in the literature [29], the components that were described before are essential to IDS. Observation is being carried out on everything that takes place on the network.

B.IDS'STAXONOMY

As its name suggests, an intrusion detection system (IDS) is able to quickly identify and notify any odd behavior by monitoring a network or other system. This is accomplished by maintaining a close check on the system. Liao et al. [31] provide a classification of intrusion detection systems that is based on four fundamental characteristics. Availability of information, instability of the system, detection method, and response time are all factors to consider." Figure 2.4 is an illustration of the different Intrusion Detection Systems (IDSs) that are stated that can be found in reference [31].

The classification of intrusion detection systems (IDSs) enables us to arrange and grasp IDS technology in accordance with its application, attack identification capabilities, data source, response methods, and system architecture. This category may be used by specialists, system designers, and security managers in order to choose the intrusion detection system (IDS) solutions that are the most appropriate for the company's needs, network size, threat models, and performance requirements. The vocabulary that is used to describe intrusion detection systems has also changed in tandem with the growing sophistication of cyber threats. The consequence of this is the development of detection methods that are both more complex and multi-modalities.

One of the most important aspects of IDS classification is the categorization of intrusion detection systems (IDSs) according to their deployment location, which in turn shows the system component that the IDS is a part of. Host-based intrusion detection systems, also known as HIDS, are responsible for monitoring system calls, application logs, changes to file systems, and user activity on each individual host or server location. Through the use of HIDS, it is possible to identify attempts at insider threats, elevated privileges, and unwanted file alterations all at the same time. Additionally, they provide easy access to a wide variety of local activities. On the other hand, they need to be installed and maintained on every

system that is being monitored, which may result in an increase in costs in settings that are on a big scale. On the other hand, network intrusion detection systems (NIDSs) are strategically positioned at crucial network nodes such ports, switches, and routers in order to continually monitor all network traffic. The contents and headers of packets are analyzed by network intrusion detection systems (NIDS) in order to identify vulnerabilities that are network-based, assaults that denial of service, and odd communication patterns. It is possible for network intrusion detection systems (NIDS) to struggle with protected data and fail to offer a full picture of host-level activities, despite the fact that they have the capability to defend several systems at the same time. The integration of host-based and network-based methodologies is what hybrid intrusion detection systems (IDS) do in order to give complete and comprehensive protection. This is accomplished by increasing the beneficial qualities of each approach while simultaneously minimizing the bad aspects of each method.

LITERATURE SURVEY

In the fields of data science and information research, dimensionality reduction is a significant phenomenon that has not yet been addressed. In order to assist in dimensionality reduction in very large datasets, a great number of IDS models have been created over the course of the last several decades. KDD99, NSL KDD, and the other networks are included in this group. Numerous studies have used the FSA to enhance the effectiveness of intrusion detection systems (IDS) and circumvent issues that are connected with high-dimensional data. On the other hand, professionals continue to struggle with the management of workloads and the simplification of data [43]. The exponential development in network traffic has resulted in an increase in the number of possible threats that might potentially occur. As a consequence of this, a number of experts have implemented a variety of machine learning strategies for Intrusion Detection Systems (IDS) that are based on Finite State Automata (FSA).

By using Support Vector Machines (SVM) and neural networks (NN), Mukkamala and colleagues [44] were able to evaluate intrusion detection systems. During the course of the testing, Support Vector Machines (SVMs) demonstrated that they are both highly adaptable and efficient when working with large datasets. The amount of time that NN must dedicate to learning is going to be enormous. A approach known as the joint information technique was used by Fleuret et al. (2004) in order to ascertain which qualities were pertinent to this debate. In comparison to SVM on its own, this method performs much better when paired with a Bayes network. The vast bulk of their study has been on the total amount of time spent working [45]. In the year 2005, Chebrolu and his colleagues conducted research on intrusion detection systems, sometimes known as IDS. The investigation took use of cutting-edge technologies such as Bayes networks and reverse classification trees, among other technological advancements. Through the use of twelve essential features that were generated from their technique, they were able to effectively identify and avoid various different types of attacks. The detection rates of U2R attacks have been shown to be much higher than predicted [46]. A number of unique feature selection algorithms (FSAs), such as correlation-based feature selection (CFS) and quick CFS, were used by Chou et al. (2008) in order to solve problems that were associated with data that had several dimensions. Among the problems is the fact that the statistics are not clear, they are not definite, and they are repeating.

4. AN EMBEDDED LEARNING ALGORITHM MODEL FOR INTRUSION DETECTION SYSTEMS

Through the identification of essential components that have the potential to enhance the performance of the recognition engine, this chapter sets the framework for FST-based systems. In order to get better results, a significant variety of algorithms have been developed that make use of a technique known as recursive feature elimination (RFE).

Both the goal and the requisite qualities have been attained, and the target has been successfully completed. For the purpose of developing and assessing the techniques, we make use of the NSL KDD Dataset. By carrying out this action, you will be able to illustrate the accuracy gap that exists between picking any features and selecting the appropriate features. In this section, ensemble classifiers like GB, AB, ET, and RF classifiers are contrasted and investigated for their performance with RFE. There are also some other classifiers that are considered. These algorithms are used to organize the data into different groups. According to the findings of comparative study, the success rate of the classifier as well as its overall performance may be significantly improved by picking the relevant characteristics with great care. I would like to provide an outline of the main sources that were used in the compilation of this chapter.

Following the use of the UFS, the RFE approach, in combination with the ANOVA F-Test and the select_Percentile process, was utilized in order to ascertain to what extent the components were significant.

A. PROPOSED FRAMEWORK

A full breakdown of the processes that will be carried out in order to put the proposed system into action is shown in Figure 4.1. The majority of the time, this strategy is comprised of four key parts. The steps consist of the following: acquiring the dataset, initializing it, doing pre-processing, selecting traits, reducing them, and analyzing them. Both "Kaggle," an online

platform for data analysis and modeling, and "Sklearn," a Python program that is specifically designed for machine learning, are used in the construction of the majority of these techniques [88]. Users of Windows 11 Pro have the ability to access, share, and analyze data with the help of a powerful processor that is the Intel(R) Core(TM) i7-10700 CPU 2.90GHz. There is a maximum of 16 gigabytes of random access memory (RAM) and 4.9 gigabytes of disk space that may be made accessible to registered users.

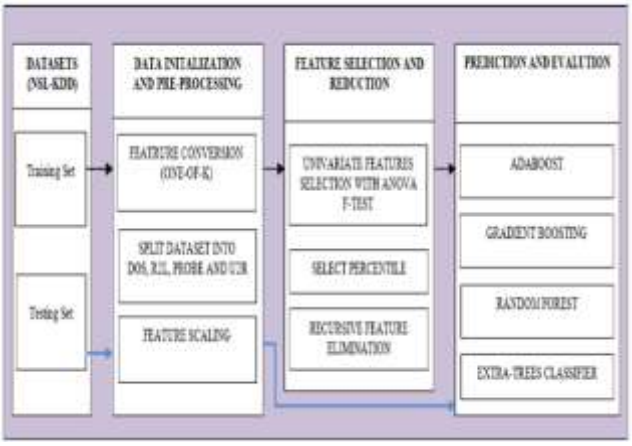

Figure 4.1: Proposed IDS Frame Work

A reduction in the number of components that make up the criteria is done in order to determine which aspects are essential. The splitting of a node in accordance with a predetermined set of values is triggered at each and every event. Table 4.6 displays the thirteen features that were retrieved from the datasets for each of the techniques. The selection of these traits was done by hand because of how important they both are.

When working with models, one of the most essential goals to strive for is to cut down on the amount of time spent working.

The outcomes of the Dos, Probe, R2L, and U2R threat cross validation (CV) tests, which were performed using RFECV and AdaBoost, respectively, are shown in Figures 4.3, 4.4, 4.5, and 4.6 in that order. On one side, we are able to see the total number of individual qualities that were chosen, and on the other side, we are able to view the CV ratings for each of those characteristics.

It is possible to see the CV that was generated by using RFECV with Gradient boosting Dos, Probe, R2L, and U2R assaults in Figures 4.7, 4.8, 4.9, and 4.10. Additionally, the CV score for each of those qualities is shown along the y-axis, while the x-axis displays the total number of features that were chosen.

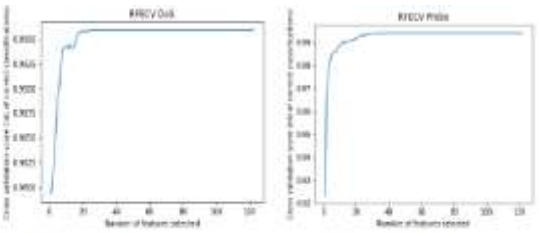

Figure 4.3 DoS RFECV with AB

Figure 4.4 PROBE RFECV with AB

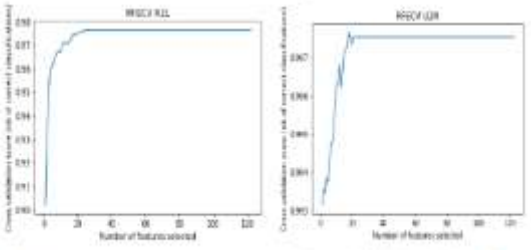

Figure 4.5 R2L RFECV with AB

Figure 4.6 U2R RFECV with AB

5. Machine Learning–Based IDS Model Using CICIDS2017 and NSL-KDD Datasets.

The purpose of this chapter is to provide you with an innovative method for using FSA to recognize notable characteristics and eliminate those that are not important. In addition to this, it validates the intrusion detection system (IDS) sensitivity of the NSL-KDD datasets to assaults.

A series of tests was performed on the engine in order to establish which of the available predictors was the most accurate. Through the use of the CICIDS2017 real-time dataset, it was feasible to assess the most effective FST and predictor. The tests that are given in this chapter indicate how the major properties of the proposed model considerably minimize the amount of processing that is required and increase the performance of the intrusion detection system (IDS). The recommended model increased accuracy by 99.94% when tested on the CICIDS2017 dataset, and it improved accuracy by 99.21% when tested on the NSL-KDD dataset. Both of these results were accomplished via testing.

A. **Proposed framework**

Because of the complexity of high traffic and the need to find a balance between a high detection rate and cheap processing costs, it is difficult to build Intrusion Detection System (IDS) models that are both effective and cost-efficient. Consequently, a classifier that is consistent with FSA is presented by this study. Its customizable and practical structure makes it feasible to reduce processing expenses while simultaneously increasing the detection rates of intrusion detection systems (IDS). It is the major goal of the system to get highly exact results while minimizing the amount of computations required. As shown in Figure 5.1, the suggested framework consists of five primary steps: the first is the collection of datasets, followed by the pre-processing of data, the FSA, the creation and evaluation of models, and finally, the analysis and selection portion. We will go into deeper depth about each step in the following paragraphs.

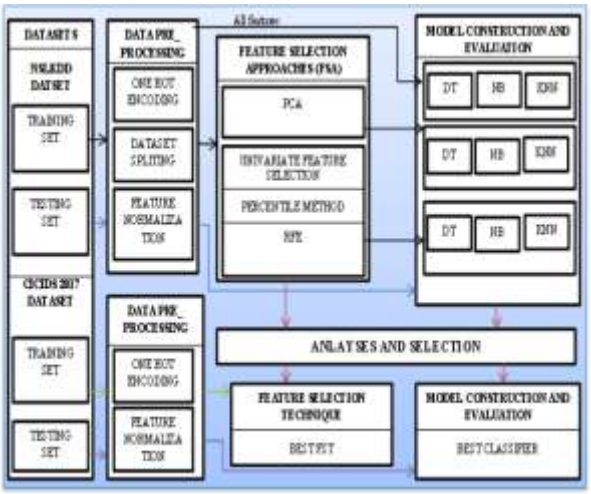

Figure 5.1: Proposed framework

CONCLUSION

In this chapter, we take a look at a number of different classifiers that integrate a number of FSTs in order to produce a powerful intrusion detection system modeling. The results of the research show that decreasing the number of datasets in IDS fulfills two goals: one is to improve the performance of the model, and the other is to accomplish another aim. expenditures related to handling are reduced. On the NSL-KDD dataset, the DT classifier that uses RFE as FST performs better than other classifiers that use FSA. The only exception to this is the U2R attack group. When it comes to accuracy, precision, recall, and the F-measure, this is absolutely correct. When contrasted with other algorithms that make use of FSA, this one functions in a distinct manner. Furthermore, the suggested FST has identified an enhanced and more condensed collection of features. This was accomplished by using ranking methodologies and information gain for the models. To do this, the FST that was given was used. According to the results of the research, thirteen essential characteristics were known to exist in the NSL-KDD dataset, whereas eight critical characteristics were found in the CICIDS 2017 dataset. The performance of the model may be improved while simultaneously reducing the amount of computational resources that are required by reducing the number of features that are utilized in the model. Through the use of the Realtime dataset (CICIDS2017), evaluations were carried out to assess the RFE+DT model's recall, G-means, precision, sensitivity, F-measure, accuracy, training time, and testing time. By contrasting it with other well-known models that have been addressed in the past, it was shown that the model that was developed is more advantageous and productive. A number of studies have shown that the use of Decision Trees (DT) as the classification method and Recursive Feature Elimination (RFE) as the Feature Selection technique (FST) leads to an improvement in the outcomes while simultaneously minimizing the amount of computer resources that are required.

REFERENCES

[1]. Zhang, Y., Li, P., & Wang, X. (2019). Intrusion detection for IoT based on improved genetic algorithm and deep belief network. IEEE Access, 7, 31711-
[2]. Elmasry, W., Akbulut, A., & Zaim, A. H. (2020). Comparative evaluation of different classification techniques for masquerade attack detection. International Journal of Information and Computer Security, 13(2), 187-209.
[3]. Shelke, M. P. K., Sontakke, M. S., & Gawande, A. D. (2012). Intrusion detection system for cloud computing. International Journal of Scientific & Technology Research, 1(4), 67-71.
[4]. Rajput, D., & Thakkar, A. (2019). A survey on different network intrusiondetection systems and countermeasure. In Emerging Research in Computing Information, Communication and Applications: ERCICA 2018, Volume 2 (pp497-506). Springer Singapore.
[5]. Wang, C., Zhao, T., & Liu, Z. (2020). An activity theory model for dynamic evolution of attack graph based on improved least square genetic algorithm. International Journal of Information and Computer Security, 12(4), 397-415.
[6]. Larson, D. (2016). Distributed denial of service attacks–holding back the flood. Network Security, 2016(3), 5-7.
[7]. Vijayakumar, D. S., & Ganapathy, S. (2022). Multistage ensembled classifier for wireless intrusion detection system. Wireless Personal Communications, 122(1), 645-668.
[8]. Alkasassbeh, M. (2017). An empirical evaluation for the intrusion detection features based on machine learning and feature selection methods. arXiv preprint arXiv:1712.09623.
[9]. Gu, S., Cheng, R., & Jin, Y. (2018). Feature selection for high-dimensional classification using a competitive swarm optimizer. Soft Computing, 22, 811- 822.
[10]. Rao, H., Shi, X., Rodrigue, A. K., Feng, J., Xia, Y., Elhoseny, M., ... & Gu, L. (2019). Feature selection based on artificial bee colony and gradient boosting decision tree. Applied Soft Computing, 74, 634-642.
[11]. Mafarja, M., Aljarah, I., Faris, H., Hammouri, A. I., Ala'M, A. Z., & Mirjalili, S.(2019). Binary grasshopper optimisation algorithm approaches for feature selection problems. Expert Systems with Applications, 117, 267-286.
[12]. Thanh, H., & Lang, T. (2019). An approach to reduce data dimension in building effective network intrusion detection systems. EAI Endorsed Transactions on Context-aware Systems and Applications, 6(18).
[13]. Almseidin, M., Alzubi, M., Kovacs, S., & Alkasassbeh, M. (2017, September).Evaluation of machine learning algorithms for intrusion detection system.In 2017 IEEE 15th International Symposium on Intelligent Systems andInformatics (SISY) (pp. 000277-000282). IEEE.
[14]. Kok, S. H., & Abdullah, A. NZJhanjhi, and Mahadevan Supramaniam. A review of intrusion detection system using machine learning approach. International Journal of Engineering Research and Technology, ISBN 0974, 3154(12), 1.
[15]. Al-Jarrah, O. Y., Siddiqui, A., Elsalamouny, M., Yoo, P. D., Muhaidat, S., & Kim, K. (2014, June). Machine-learning-based feature selection techniques for large-scale network intrusion detection. In 2014 IEEE 34th international conference on distributed computing systems workshops (ICDCSW) (pp. 177- 181). IEEE.