International Journal of Engineering Technology and Applied Science
(ISSN: 2395 3853), Vol. 7 Issue 6 June 2021

A Survey on Intrusion Detection Systems for Detecting Networking Attacks using NSL-KDD Dataset

Ms. Swikrati Dubey[1], Prof. Chetan Gupta[2], Prof. Shivendra Dubey[3]

swikratidubey@gmail.com1, chetangupta.gupta1@gmail.com[2], Shivendra.dubey5@gmail.com[3]

M.Tech. Scholar, Dept. of CSE[1], Asst. Prof., Dept. of CSE[2,] Asst. Prof., Dept. of CSE[3]

SIRTS Bhopal[1], SIRTS Bhopal[2,] SIRTS Bhopal[3]

Abstract—Cyber-attacks are becoming more sophisticated and thus presenting increasing difficulties in reliably detecting intrusions. Failure to deter the intrusions may degrade the credibility of security services, e.g. data confidentiality, integrity, or availability. Intrusion detection is increasingly critical and difficult with the use of networking technologies and the Internet. A variety of strategies were developed for the identification of intrusions by machine learning and deep learning. This research provides taxonomy of contemporary IDS, an outline of important recent papers and an overview of the data sets used to perform the assessment. It also presents evasion methods used by attackers to avoid detection and discusses future research problems in effort to fight those techniques in order to improve the protection of computer networks. There has been a graphical analysis of different machine learning algorithms.

Keywords—Network security, cyber-attacks, intrusion detection system, Machine learning (ML) algorithms, NSLKDD.

I. INTRODUCTION

Cyber security is a significant research area, and the networks have expanded impacts on daily life. Techniques in cyber security cover primarily anti-virus applications, firewalls, and IDSs. This defends networks from internal and external threats. One of these is an IDS, which theaters crucial role in protecting cyber security by tracking the status of the network's software and hardware. In 1980, they proposed the first intrusion detection system [1]. Since then there have been several mature IDS goods. Many IDSs, however, still have a high false alarm rate which produces many warnings for low – anti situations, that increases the workload on security analysts or may lead to serious harmful assaults being ignored. Numerous researchers have therefore intensive on developing IDSs with higher levels of detection or lower false alarm rates. The fact that they have no unknown attacks is another issue with current IDSs. Owing to increasingly evolving network environments, threat types and new attacks are continuously emerging. Thus, IDSs that can detect unknown attacks must be developed.

Researchers started to focus on developing IDSs using ML methods to discuss the above problems. ML is a kind of artificial intelligence technology that can find information from large datasets automatically [2]. When more data is given &ML models can be extended enough to identify attack variations and new threats, machine-based IDSs can achieve satisfying levels. Besides that, IDSs focused on machine learning do not largely rely on field knowledge; they are therefore simple to plan and develop.

Intrusion helps researchers, in the context of IDS, to unlawfully or unauthorized access to data around computer systems or damage systems operation. An IDS is an application for computer security to notice wide variety of security violations, beginning with attempted interruptions by external entities, and system intrusion and insider abuse [3]. The IDS's main function is to track hosts and networks, analyze computer systems' behaviors, generate a warning, or respond to unusual behavior. Since similar hosts or networks are monitored, IDSs are usually introduced near secured network nodes (e.g. transitioning in key network segments).

IDS classification methods are used in 2 kinds: a method based on identification as well as a method based on data source. IDSs can be decomposed into misuse detection & anomaly detection among detection-based approaches. IDSs may be separated into host-based methods as well as network-based methods among the data source methods [4]. This survey incorporates these 2 types of IDS classification methods, takes as a primary consideration the data source, and takes a secondary classification feature the detection system.

Classification of IDS:

A. Host-based IDS (HIDS)

HIDS or host-installed software applications to be monitored. The operating system is tracking and the data is entered in log files and alerts are activated. Only the workstations on which the agents are located can be observing. Host-based IDS services are used to track vital server attack attempts. Host-based IDS looks at the local system intrusion sign. An audit trail is the trust of the host-based system. The details help IDS to detect subtle trends of misuse not seen at a higher abstraction level.

B. Network-based IDS (NIDS)

The NIDS consists typically of a network computer (or sensor) with a promiscuous Network Interface Card (NIC) as well as a

different management interface. The IDS is located along with a network or border and controls all traffic in that area. These systems gather data from the network itself not from each specific host. The network attacks are reviewed by the NIDS as packets pass across the network. The openness of monitors eliminates the risk that an opponent will find the display and cancel its capacities without effort. Network Node IDS (NNIDS) agents are installed on each host that is secured in the network.

C. Application-based IDS

IDS based on device are a special HIDS subset that analyzes activities within a software application. The application's transaction log file is the most common information source for application-based IDS.

II. LITERATURE REVIEW

Various researchers have been working on this concept of networking security due to the increasing attacks in the network.

S. S. Ahmadi et al. [2019]To create successful IDS, this study centered on defining key attributes Using three common methods for choosing the most appropriate features for IDS, a majority voting system, correlation-based feature selection, Information Gain, or Chi-square is proposed. Decision tree classifier is used in the development of an intrusion detection method with reduced feature sets. The findings show that chosen reduced attributes from the current role selection method increase the output to create an IDS system that is computer-efficient. [5].

L. Hakim et al. [2019] the performance as well as the training data used by an IDS depend on its algorithm. Important training data features can reduce IDS detection efficiency and accuracy. This thesis would investigate the effect on the intrusion detection system by using feature selection. In J48, Random Tree, Naïve Bayes & KNN algo, knowledge gain, Gains Ration, Chi-squared, and Relief methods of selecting the effect will be studied. The results show that the range of features can dramatically boost IDS efficiency, although the inaccuracy is reduced slightly [6].

K. A. Taher et al. [2019]This study finds that wrapper feature selection Artificial Neural Network (ANN)-based machine learning outperforms the vector machine support technique (SVM) when network traffic classification. NSL-KDD is used for the assessment of performance to identify network traffic using machine learning techniques supervised by the SVM and ANN method. The comparative analysis indicates that the model proposed is efficient about the success rate of intrusion detection than other current models [7].

M. M. Sakr et al. [2019] several techniques for feature selection are introduced in this paper to improve NIDS efficiency. Sensors or wrappers (Genetic Algorithm (GA), Artificial Bee Colony (ABC) & Particle Swarm Optimization (PSO) are the types of the selected techniques utilized Information Gain (IG), Principal Component Analysis (PCA) and Correlation Function Collection (CFS)). SVM for a description of network connections. To develop and test NIDS, the NSL-KDD network traffic benchmark is chosen. The evaluation findings showed that wrapper approaches obtained higher categorized accuracy, identification rate, real positive rates & low false-positive rates than filter methods for NIDS. Associated with other linked NIDS, their ABC-NIDS has been seen to be the best value in their system. [8].

S. Sun et al. [2018] in the paper, a lightning attachment procedure optimization algorithm (LAPO) & SVM for intrusion detection are proposed for the wrapper feature selection system. LAPO is a newly suggested, versatile searchable natural- inspired algorithm. The famous KDD Cup 99 dataset is applied to test the efficiency of the proposed system. Experimental results suggest a greater efficiency and precision when looking for the optimum function subset compared with the GA or PSO. [9].

H. M. Answer et al. [2018] this study contains a feature selection system with numerous classification devices for effective network anomaly detection. The approach uses to filter and wrapper filtering methodologies to incorporate various methods. The goal of this frame is to pick the least no. of features to ensure optimum precision. The data set UNSW-NB15 for evaluation of the proposed framework is included in experimental findings. The findings revealed that the accuracy of 88 percent is achieved with 18 features based on one of the filters and J48 as a classifier [10].

H. Li et al. [2018] in this paper suggest a model for hybrid FS based on optimization of random forest or particle swarm, using both an independent calculation & an educational algorithm. It selects the best subsets for a given cardinality using the independent measure & uses a learning algorithm to pick the final best subset from among the best subsets in cardinalities. 1999 KDD dataset was used as an occasion for testing TPR & FPR of proposed model & for comparing them to CFS & SVM algorithms. [11].

J. Ku et al. [2017] ID is now a promising area in security research. IDSs have evolved as a vital module as well as an important network security technology, & IDS can automatically ID by someone who is not allowed into the current computer system. ELM is a research field specialized in identifying potential threats or intrusions. In this text, they suggest an improved learning algorithm called self-adaptive differential evolution ELM (SADE-ELM) for the classification and detection of intrusions. We use ELM, DE-

ELM procedures, in our classifications, to compare our approaches. The SADE-ELM method suggested indicates that the identification accuracy of classification cases is better than the proposed one. [12].

The problem of suggestion mining as presented in the SemEval 2019, is posed as a binary classification problem and could be formally stated as: For a dataset D of sentences named, the objective is to learn a classification/prediction function, which can predict a mark l for a sentence s, where l ∈ {suggestion, non-suggestion}.

1) Differential evolution feature selection

Differential evolution (DE) is one of the evolutionary algorithm approaches in which the features are searched on basis of an ant colony. Such as other optimization methods [12], [13] DE offers the advantages, which are simple yet efficient. The DE compensation is several: 1) ability to handle no differentiable, non-linear as well as multimodal functions of costs; 2) parallelizability to handle analysis cost functions; 3) ease of use; As GA, DE uses the same mutation, crossover or selection criteria. The efficiency of DE depends on the handling of the target vector or vector to acquire a test vector during the searching process.

2) Whale Optimization Algorithm (WOA)

Mirjalili or Lewis introduced the WOA in 201634. This algorithm contains two major phases: first-stage encircling or first spiral updating. A random search for a target (exploration stage) will be carried out in the second process. At the onset, whales are given arbitrary solutions as well as the min or max value of an objective function is considered to address the best value based on the situation. Each objective function search agent is then determined. Every search agent changes its position based on the best solution or random search agent for each iteration.

3) Extreme Learning Machines (ELM)

Using ELM to train a single hidden layer NN (SLFNs). In ELM, hidden nodes are initiated randomly or fixed without iterative tuning. In comparison, hidden nodes of ELM do not even need to be like a neuron. The free parameter it requires to learn is links (or weights) of the output layer to the hidden layer. ELM is thus built as a linear model in a parameter that essentially goals to solve a linear system. Directly opposed to standard FNN methods of instruction, ELM is much more effective and aims to attain the worldwide optimum. ELM has been shown to preserve the universal approximation potential of SLFNs even though it deals with hidden nodes randomly generated.

4) Simulated Annealing (SA)

SA, inspired by the method of metallurgy, is a met heuristic solution. It is an easy method of maximizing substance heating and cooling to intensify the crystal scale. The energy is reduced to eliminate defects in metal structures by room temperature. As a controlling factor and an internal energy feature, the SA technique utilizes its temperature progress. The simulated rectification begins with a primary S solution as well as an updated S′ solution. If the fitness function F(S*) values are smaller than F(S) the solution for this procedure is generated.

$$P_b = \exp \qquad (1)$$

The higher S* fitness value is recognized as described in Eq. [6]. This policy makes it possible to eliminate interaction with local optima via the search process. F(S*) is the neighboring solution's fitness function, so F(S) is the present solution's fitness function. The control parameter is defined by Temperature Tm. The balance is achieved based on the sequence of moves &the parameter of temperature control is determined on basis of the cooling rate. Tm parameter control affects global search efficiency. Simulated anneal mechanism has a better probability if the temperature gets a high initial value. The SA procedure will stop after a series of temperature decrease if no improvements are created. If the initial temperature is low &the calculation time is shorter, the possibility of searching global solutions is further reduced.

$$T_m = \delta^k + T_o + T_{fn} \qquad (2)$$

In cases where δk is the decreasing T_m, $0<\delta<1$, k is the number of stints provided by the neighboring solution; To is the initial temperature value &T_{fn} is the final temperature value. The procedure for SA is described below algorithm.

Table 1: Literature Survey of IDS

S. No.	Year of Publication	Author	Algorithm	Accuracy
1	2020	Sugianela, Y., & Ahmad, T. [13]	Random Forest classifier	99.36%
2	2020	Faezah Hamad Almasoudy, WathiqLaftah Al-Yaseen, Ali KadhumIdrees [14]	DE feature selection with ELM classifier	80.15 %
3	2020	Matel, E. C., Sison, A. M., & Medina, R. P. [15]	Genetic Algorithm with improved feature selection (GA-IFS)	80.47%
4	2019	S. S. Ahmadi, S. Rashad & H. Elgazzar	decision tree, information gain, Chi-	79.96%, 79.91%, 79.91%,

		[16]	square method, trial and error method	75.30%
5	2018	H. M. Anwer, M. Farouk & A. Abdel-Hamid [17]	J48 classifier	88%
6	2018	Chen, F., Ye, Z., Wang, C., Yan, L., & Wang, R. [18]	TSA-KNN	80.02%
7	2017	Jabbar, M. A., Aluvalu, R., & Satyanarayana Reddy, S. S. [19]	Bayesian network	99.9%
8	2017	Shao-Bo, D. [20]	Intrusion Feature Selection Method Based on Neighborhood Distance(IFSMND)	96.9%
9	2017	X. Zhang, P. Zhu, J. Tian & J. Zhang [21]	Laplacian Support Vector Machine (LapSVM)	97.8%

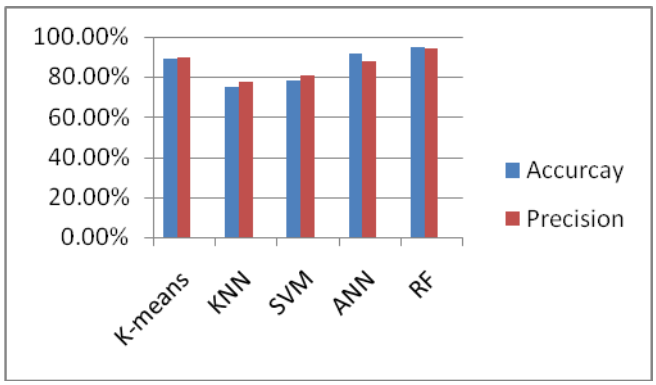

Figure 1 represents the comparison graph of Machine learning algorithms where K-means, KNN, SVM, ANN.

Data Description

There are several intrusion detection datasets but this study is based on a particular dataset which is the KDD'99 dataset. As a newly modified version of the original KDDCup'99 dataset, the NSL-KDD data set was proposed in 2009. The benefits and difficulties of KDD-Cup 99, on the first hand, have been maintained by NSL-KDD. Only in reducing redundancies, rationalization of numbers of cases, or preservation of the variability of the selected samples, did the study discuss those drawbacks inherited from the original results. Make sure that the dataset NSL-KDD is compiled to maximize the predictive difficulty which makes it outstanding. The first dataset was evaluated using multiple benchmark classifiers to group the records in five complexity ranges, or each case was annotated as various predictions [12]. The number of selected data is inversely proportional to the record percentages of the original KDDCup99 dataset for each complex level group.

Every record is listed as usual or abnormal if the abnormal one is 22 attacks in the training set & 39 attacks in the test set. [21]:

- **DOS:** Resources have been allocated by more device demands to avoid users' availability.
- **Probe:** Check by network scanning for information about the target host.
- **User to Root (U2R):** Request unauthorized access to the controlling account by a devaluation of the device details by the password.
- **Remote to User (R2U):** Legal user access to the device.

III. PROBLEM DOMAIN

As per the previous researches and respective research paper, following are the various problem domains which I will try to resolve and improve in my research work.

- Less Accuracy
- Detection Rate is low
- Precision percentage is low
- F-Score is less
- False Alarm Rate is less

IV. PROPOSE WORK

After reviewing various Research paper and research problem, I am trying to propose a Hybrid Intrusion Detection System, which will be based on Machine Learning Approach using either supervised learning or Un Supervised Learning techniques.

Additionally, I will also use a Data Mining Algorithm for data classification. And, this Hybrid combination will produce better and enhanced result as compared to previous work, and I will try to evaluate the proposed work by checking the following parameters:

- Accuracy:
- Detection Rate
- Precision
- F-Score
- False Alarm Rate

V. CONCLUSION

By utilizing innovative methods as well as social networking approaches, cybercriminals attack computer users. Some cybercriminals are more sophisticated as well as motivated. Cybercriminals have shown their ability to hide their names, hide their communication, separate their identities from illegal profits, or use resilient infrastructure. Thus, computers with advanced IDSs capable of detecting modern malware are increasingly necessary to secure. A complete overview of the strengths and limitations of modern IDS research is important for developing or building such IDS systems. We also presented a detailed survey of approaches, types, or technologies for the intrusion detection system with its benefits and constraints.

REFERENCES

1. Anderson, J.P. Computer Security Threat Monitoring and Surveillance; Technical Report; James P. Anderson Company: Philadelphia, PA, USA, 1980.
2. Michie, D.; Spiegelhalter, D.J.; Taylor, C. Machine Learning, Neurall, and Statistical Classification; Ellis Horwood Series in Artificial Intelligence: New York, NY, USA, 1994; Volume 13.
3. Denning, D.E. An intrusion-detection model. IEEE Trans. Softw. Eng. 1987, 222–232.
4. Heberlein, L.T.; Dias, G.V.; Levitt, K.N.; Mukherjee, B.; Wood, J.; Wolber, D. A network security monitor. In Proceedings of the 1990 IEEE Computer Society Symposium on Research in Security and Privacy, Oakland, CA, USA, 7–9 May 1990; pp. 296–304.
5. S. S. Ahmadi, S. Rashad and H. Elgazzar, "Efficient Feature Selection for Intrusion Detection Systems," 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York City, NY, USA, 2019, pp. 1029-1034, DOI: 10.1109/UEMCON47517.2019.8992960.
6. L. Hakim, R. Fatma, and Novriandi, "Influence Analysis of Feature Selection to Network Intrusion Detection System Performance Using NSL-KDD Dataset," 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), Jember, Indonesia, 2019, pp. 217-220, DOI: 10.1109/ICOMITEE.2019.8920961.
7. K. A. Taher, B. Mohammed Yasin Jisan and M. M. Rahman, "Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection," 2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST), Dhaka, Bangladesh, 2019, pp. 643-646, DOI: 10.1109/ICREST.2019.8644161.
8. M. M. Sakr, M. A. Tawfeeq and A. B. El-Sisi, "Filter Versus Wrapper Feature Selection for Network Intrusion Detection System," 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt, 2019, pp. 209-214, DOI: 10.1109/ICICIS46948.2019.9014797.
9. S. Sun, Z. Ye, L. Yan, J. Su, and R. Wang, "Wrapper Feature Selection Based on Lightning Attachment Procedure Optimization and Support Vector Machine for Intrusion Detection," 2018 IEEE 4th International Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS-SWS), Lviv, 2018, pp. 41-46, DOI: 10.1109/IDAACS-SWS.2018.8525742.
10. H. M. Anwer, M. Farouk, and A. Abdel-Hamid, "A framework for efficient network anomaly intrusion detection with features selection," 2018 9th International Conference on Information and Communication Systems (ICICS), Irbid, 2018, pp. 157-162, DOI: 10.1109/IACS.2018.8355459.
11. H. Li, W. Guo, G. Wu, and Y. Li, "An RF-PSO Based Hybrid Feature Selection Model in Intrusion Detection System," 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), Guangzhou, 2018, pp. 795-802, DOI: 10.1109/DSC.2018.00128.
12. J. Ku, B. Zheng, and D. Yun, "Intrusion Detection Based on Self-Adaptive Differential Evolutionary Extreme Learning Machine," 2017 International Conference on Computer Network, Electronic and Automation (ICCNEA), Xi'an, 2017, pp. 94-100, DOI: 10.1109/ICCNEA.2017.57.
13. Sugianela, Y., & Ahmad, T. (2020). Pearson Correlation Attribute Evaluation-based Feature Selection for Intrusion Detection System. 2020 International Conference on Smart Technology and Applications (ICoSTA).
14. Almasoudy, F. H., Al-Yaseen, W. L., & Idrees, A. K. (2020). Differential Evolution Wrapper Feature Selection for Intrusion Detection System. Procedia Computer Science, 167, 1230–1239.
15. Elmer C., Ariel M. Sison Ruji P. Medina Matel "Implementation of GA-IFS-based Network Intrusion Detection System: A comparative analysis" ICSET'20: 2020 The 4th International Conference on E-Society, E-Education and E-Technology.
16. S. S. Ahmadi, S. Rashad and H. Elgazzar, "Efficient Feature Selection for Intrusion Detection Systems," 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York City, NY, USA, 2019, pp. 1029-1034, DOI: 10.1109/UEMCON47517.2019.8992960.
17. H. M. Anwer, M. Farouk, and A. Abdel-Hamid, "A framework for efficient network anomaly intrusion detection with features selection," 2018 9th International Conference on Information and Communication Systems (ICICS), Irbid, 2018, pp. 157-162, DOI: 10.1109/IACS.2018.8355459.
18. 6. F. Chen, Z. Ye, C. Wang, L. Yan, and R. Wang, "A Feature Selection Approach for Network Intrusion Detection Based on Tree-Seed Algorithm and K-Nearest Neighbor," 2018 IEEE 4th International Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS-SWS), Lviv, 2018, pp. 68-72.
19. M. A. Jabbar, R. Aluvalu and S. S. Satyanarayana Reddy, "Intrusion Detection System Using Bayesian Network and Feature Subset Selection," 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Coimbatore, 2017, pp. 1-5.
20. D. Shao-Bo, "Intrusion Feature Selection Method Based on Neighborhood Distance," 2017 International Conference on Computer Systems, Electronics, and Control (ICCSEC), Dalian, 2017, pp. 748-751, DOI: 10.1109/ICCSEC.2017.8446849.
21. X. Zhang, P. Zhu, J. Tian, and J. Zhang, "An effective semi-supervised model for intrusion detection using feature selection based LapSVM," 2017 International Conference on Computer, Information and Telecommunication Systems (CITS), Dalian, 2017, pp. 283-286.