

Aditya Harbola, Priti Dimri, Deepti Negi

Abstract: At present networking technologies has provided a better medium for people to communicate and exchange information on the internet. This is the reason in the last ten years the number of internet users has increased exponentially. The high-end use of network technology and the internet has also presented many security problems. Many intrusion detection techniques are proposed in combination with KDD99, NSL-KDD datasets. But there are some limitations of available datasets. Intrusion detection using machine learning algorithms makes the detection system more accurate and fast. So in this paper, a new hybrid approach of machine learning combining feature selection and classification algorithms is presented. The model is examined with the UNSW NB15 intrusion dataset. The proposed model has achieved better accuracy rate and attack detection also improved while the false attack rate is reduced. The model is also successful to accurately classify rare cyber attacks like worms, backdoor, and

Keywords: Intrusion Detection, Feature Selection, machine learning, UNSW NB15.

I. INTRODUCTION

To prevent users from network attacks many traditional techniques are used such as Firewall, Authentication, Authorization, antiviruses, etc. But all these techniques may be compromised due to loopholes like compromised password, intruders and backdoor attacks. Network attacks are dynamic n nature in a sense that every day the nature of the attack changes. For changing behavior of the attacks Intrusion Detection Systems (IDSs) [1] is the best solution because it can be used with data mining techniques that can identify or predict a potential attack before it could damage the network. Generally, we come across two versions of NIDS models, misuse and anomaly detection models [2]. Misuse intrusion detection models are static in nature because it has a fixed matching pattern designed from known attack database. Misuse technique will try to match the unknown pattern with the known attack pattern and then try to detect the intrusion. In anomaly detection models the focus is on identifying the behaviors and activities which are not normal for the network.

Revised Manuscript Received on April 30, 2020.

* Correspondence Author

Aditya Harbola*, School of computing, Graphic Era Hill University, Dehradun, India. Email: adityaharbola@gmail.com

Priti Dimri, Computer science and applications, GBPEC, Ghurdouri, Pauri, Pauri, India. Email: pdimri1@gmail.com

Deepti Negi, d School of computing, Graphic Era Hill University, Dehradun, India. Email: deeptine@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

If abnormal behavior is sensed by the IDS it will raise an alarm for the intrusion. The difference between both models is that novel attacks cannot be handled with misuse models. Misuse models are very beneficial for known attacks but at present when attacks are changing their signature, each day anomaly detection models are much beneficial. The limitation of Anomaly detection is that if the model is not well trained it will lead to a false alarm. To handle novel attacks a new intrusion detection system using an online approach is needed which will be a combination of machine learning and data mining techniques. Many studies have demonstrated methods to minimize intrusion detection time and to improve accuracy. In online intrusion detection, there are few limitations in the existing studies. In the proposed models, the time to build the model is not considered. In most of the approaches, focus is not given on feature representation of the connections, which can improve intrusion detection performance.

In section 2, related work for network anomaly detection is discussed. In section 3 dataset, tools and algorithms are discussed. A feature selection model and classification techniques is covered in section 4. Evaluation of models and the obtained results are covered in section 5. In section 6 result analysis is done. The conclusion of the work is covered in section 7.

II. RELATED WORK

Many techniques and algorithms have been proposed for intrusion detection. Each technique offers some improvement in efficiency in intrusion detection rate and anomaly detection. In this section, the related work in this area is briefly resented. (S. Peddabachigari et al, 2007) has proposed a hybrid approach which combines two classifiers, Decision tree and Support vector machine. They called it DT-SVM. The idea of combining two classifiers and designing a new hybrid classifier leads to increase detection accuracy. The DT-SVM approach provided better accuracy. S. Singh et al. ,2011 designed and proposed an algorithm based on SVM and Gaussian kernel. This scalable algorithm improved SVM performance when applied on KDD-99 intrusion data set. In this work, accuracy improved and the training time of the model is decreased. R. Chitrakar et al.,2012 proposed a combination of two methods k-medoids and SVM and stated that this approach was better than a combination of k-Medoids and Naive Bayes in terms of detection rate and accuracy. (AP. Muniyandi et al., 2012) proposed a machine learning technique combining K-means and C4.5 algorithms. They made use us unsupervised learning with K-means to train a certain percentage of the training data.



Journal Website: www.ijitee.org

Then supervised learning was done with the C4.5 algorithm. This approach achieved more accuracy to detect anomalies. (VD. Katkar et al.2013) proposed a multi classifier approach to achieve high accuracy. The main focus of the work was to select the best feature of multi classifiers. The result of the work shows that the attack detection rate improved. [2] proposed a mixed attack detection method which is the integration of the misuse and anomaly detection model. They combined c4.5 and SVM and experiments were performed on NSL-KDD dataset.

C4.5 used to train the dataset and SVM used for classify the normal and abnormal behavior. (Y. Canbay et al.,2015) provides a hybrid approach using k-Nearest neighbor and genetic algorithm. In this work genetic algorithm selects k neighbors of the sample and k-NN classifies these samples. The experiment performed on KDD-cup 99 dataset shows improvement over the simple k-NN classification method. [3] work provides a mixed approach combining sequential minimal optimization and k-means algorithm. The hybrid approach of clustering and classification algorithm applied on NSL-KDD dataset. Using clustering method training dataset dimension were reduced.

The results show improvement of the anomaly detection rates while reducing false alarms. A. Harbola et al.,2014 proposed a mixed approach of reducing features using feature selection approach and then applying various machine algorithms. The results show an improvement in detection rate when only important features were used.

Testing the accuracy of any NIDS needs a dataset which contains updates normal and attack activities. The most general datasets KDD99 and NSL-KDD has the following limitations.

- a) These datasets do not have modern attacks which change their signature over time [1].
- b) Both datasets are not meeting traffic and networking requirements of the present time because they were created years ago [4].
- c) The testing and training datasets have some uncommon attacks which results in fault in intrusion detection.

So a new dataset for network intrusion attacks released in 2015 [5]. The dataset is freely available online for research purposes. UNSW-NB15 dataset is created by the IXIA PerfectStorm tool [6]. The major highlights of the dataset are:

- a) It contains nine categories of attacks.
- It contains realistic network traffic data captured over some time.
- c) It contains 49 features for each type of attack.
- d) The packet header of each data contains in-depth properties of the network traffic.

So we have used the UNSW-NB15 dataset and it is explained in the next section.

III. DATA SET, TOOL AND ALGORITHMS USED

a) UNSW-NB15 Data Set: The UNSW-NB15 dataset has two million records and 540,044 records are maintained in four CSV files named, UNSW-NB15_1.csv, UNSW-NB15_1.csv, UNSW-NB15_1.csv, and UNSW-NB15_1.csv. The distribution of the attacks is shown in table 1. The training and testing dataset has all types of attacks.

The dataset has nine different attack types as follows

- 1) Fuzzers: These type of attacks sends a huge amount of data the target machine so that machines are crashed due to inability of handling the data.
- 2) Analysis: These attacks are to analyze the network such as port, spam files or loopholes scanning. These attacks do not harm the system but scans the system.
- 3) Backdoor: In these types of attacks the attacker finds a backdoor by using some legitimate system. Then they install a program in the system to gain remote access.
- 4) Denial of service: The network resources are utilized by many illegitimate requests and the legitimate users are denied network access. These attacks are not easy to recognize.
- 5) Exploits: These type of attacks compromise the limitations or vulnerabilities of the operating system or system software's
- 6) Generic: This type of attack is done on message ciphers. It is also called a collision attack because it depends upon collision on the secret key.
- 7) Reconnaissance: These attacks are based on the network information available publically. Social media searches have made these attacks more frequent.
- 8) Shellcode: A small program is injected with the applications to gain access to the target machine. This is a subclass of exploits attack.
- 9) Worms: To convert a target machine into bots/zombies worms attacks are used. These are program which spreads through network and infects the network. Distributed attacks are done using worms. Table 1 describes the distribution of the different category normal/attacks) in the dataset

10) Table 1: Distribution of the different category of traffic(Normal/attack) in the dataset

traine	tramc(Normal/attack) in the dataset					
Category	Training set	Testing set				
Normal	56,000	37,000				
Analysis	2,000	677				
Backdoor	1,746	583				
DoS	12,264	4089				
Exploits	33,393	11,132				
Fuzzers	18,184	6,062				
Generic	40,000	18,871				
Reconnaissance	10,491	3,496				
Shellcode	1,133	378				
Worms	130	44				
Total Records	175,341	82,332				

b) WEKA tool: It is a tool for machine learning which helps in data clustering, regression, processing, classification, association rules, and visualization. WEKA comes with GNU public license and written in java [7]. The algorithms can be directly applied to the datasets or can be applied through a java code. The WEKA tool can be summarized in following figure 1.



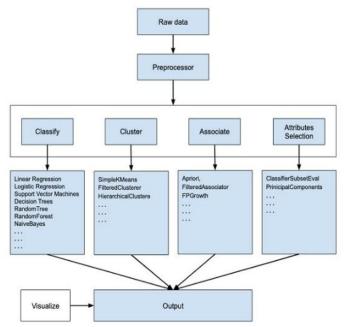


Fig 1. Various working areas of the WEKA tool.

- c) Feature Selection: Intrusion detection in real-time needs faster processing of the data. The network data has many features and to implement machine learning efficiently, only efficient features must be used. Feature selection is a crucial process inefficient intrusion detection [4]. Using only the relevant features provides more accurate predictability [8]. There are many traditional methods for feature selection i.e. information gain, correlation coefficients, subset selection, etc. We have proposed to use a hybrid method of clustering algorithms and feature selection. The hybrid method includes K-means and correlation-based feature selection methods.
- **c.1) K-Means Algorithm:** It is an iteration based clustering algorithm which tries to partition of the datasets into distinct non-overlapping groups [3]. It takes several clusters (k) and dataset as input and determines the distance of each object to the centroid. K-Means is suitable for large datasets because of its complexity which is O(Objects*clusters*iterations). The algorithm is as follows
 - i. Assign number of clusters k.
 - ii. Shuffle the dataset and initialize the centroids by selecting k data points.
 - iii. Iterate till centroid does not change
 - iv. Compute the addition of the squared distance
 - v. Assign the calculated data points to the nearest centroids
 - vi. Compute the centroid of the cluster by taking an average of the all data points which belong to each cluster.
- **c.2) Correlation feature Evaluation:** It evaluates the value of a feature by calculating the correlation (Pearson's) between it and the class. We will use correlation feature evaluation with a ranker algorithm to rank the features from most significant to least significant features[9].
- d) Classification Algorithms: Classification is a technique to label the dataset in desired categories. We will use ZeroR, IBK, and random forest classifiers. WEKA provides these classification algorithms which can be configured according to our needs. We will use these algorithms to find the confusion matrix of the algorithms with and without feature selection. Using this confusion matrix (Figure 2) we will calculate parameters; accuracy, FADR, ADR which will be used to examine if feature selection has improved the intrusion detection.

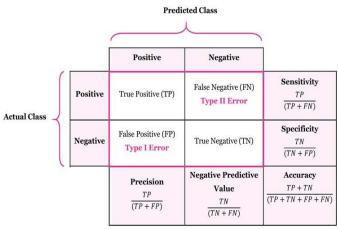


Fig 2: Confusion Matrix and classes

IV. PROPOSED MODEL

We are proposing a new model for efficient feature selection and intrusion detection. The model is shown in figure 3. This approach is as follows.

- 1) We first select the features of UNSW-NB15 datasets based on the following steps.
- step 1: Process the data set with a clustering algorithm
- step 2: Select the significant features of each attack family dataset from the cluster
- step 3: perform correlation feature selection with ranker algorithm on the dataset to select the best features.

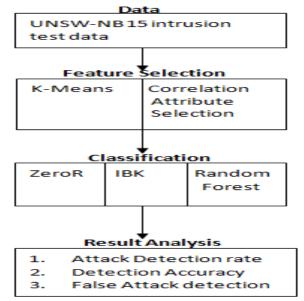


Fig. 3: Proposed model of efficient intrusion detection
2) Test the dataset with newly selected best features for each attack type

- Step 1: Analyze the dataset with all features and with selected features using various machine learning algorithms
- Step 2: Analyze the results and accuracy of machine learning algorithms.
- Step 3: Select the best performing algorithm.

V. EXPERIMENTS AND RESULTS

We have selected the half of the records of the UNSW-NB15 test data set. The selection is random and the selected records were used for the model.

After uploading the filtered data set to WEKA clustering is done using the K-means algorithm. Figure 3 shows the configuration of the algorithm used.

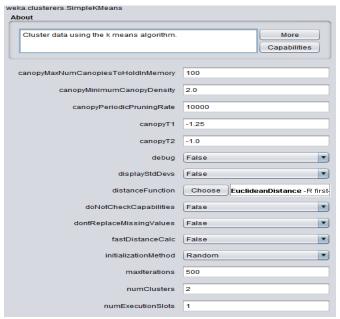


Figure 3. K-means setup with k=2

The following tables show the results of the k-means algorithm for clusters, k=2. From table 2, the label value of cluster 1 is 0.823.

Table 2: K-means with k=2

Feature	Cluster 0	Cluster 1
Protocol	TCP	UDP
Service	DNS	DNS
State	FIN	INT
Label	.5107	0.823
Sbyte	18475.5887	774.147
Dbyte	32683.4553	50.3722
Dinpkt	161.7228	26.6169
ct_srv_src	4.0031	13.7507
ct_state_ttl	0.723	1.7912
ct_dst_ltm	2.5664	9.2338
ct_src_dport_ltm	1.4007	8.7212
ct_dst_sport_ltm	1.06	6.8428
ct_dst_src_ltm	2.8401	13.6656

Table 3 shows the results of the K-means with k=8. Based on the results of K-means algorithm for k=2,k=8 clusters the highest value of the features to occur in attack data are as follows:

Protocol, Service, state, sload, sloss, is_sm_ips_port, , ct_state_ttl, ct_src_dport_ltm,

After clustering the data we performed correlation feature selection with search method of ranker. The configuration of the correlation feature selection algorithm in WEKA is shown in figure 4. Figure 5 shows the configuration of the ranker search algorithm

Table 3: K-means clustering with k=8

The state of the s								
Feature	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
protocol	tcp	udp	udp	tcp	tcp	unas	udp	Udp
service	dns	dns	-	http	-	-	dns	-
state	CON	INT	FIN	FIN	FIN	INT	INT	INT
sloss	0	0	14.39	7.22	7.29	0	0	0.16
sload	7354033	942958	100841848	1362769	668812	154439673	96448547	273390341
dload	601500	13	52702	774021	3721050	137	0	173
attack_cat	Normal	Generic	Exploits	Normal	Normal	Exploits	Generic	Normal

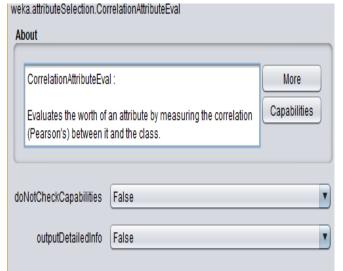


Figure 4: Correlation feature selection algorithm configuration in WEKA



Figure 5: Ranker algorithm configuration in WEKA The result of the correlation feature selection algorithm shows that the following features were selected: Protocol, Service, Sloss, Dloss, Sttl, Dttl, ct_state_ttl,ct_src_dport_ltm



Journal Website: www.ijitee.org



Table 4 presents all the selected features for each type of attack. the table gives us a summary of the most significant feature needed to identify a particular type of attack.

Table 4: Most significant attributes for each attack

Attack	Feature
Fuzzers	5,7,10,11,14,16
Analysis	5,6,7,10,11,13,14,15,36
BackDoor	4,5,6,7,10,14,16,25,33,34,35,36,37
DOS	5,6,7,10,12,14,16,17,19,20,24,28,34,36,37,41,4
	2
Exploits	6,7,10,11,12,13,16,18,1920,24,28,34,37,41,42
Generic	5,6,7,10,11,23,33,34,36,37,41
Reconnaisance	6,7,10,11,14,41,42,43,44,47
ShellCode	7,10,11,14,41,42,44
Worms	5,10,13,14,25,38,41,44

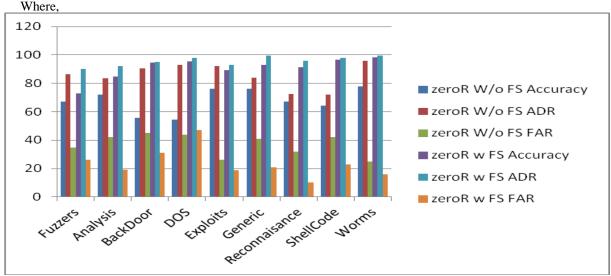
After selecting the most significant features for each attack category three classification algorithms were used: ZeroR, IBK, Random Forest. These classification algorithms are available in the WEKA tool. After setting the desired configuration the results were compared on the following parameters.

- 1. Attack Detection rate: ADR=TP/(TP+FN)
- 2. Detection Accuracy: (TP+TN)/total observations
- 3. False Attack detection: (FP+FN)/ total observations

VI. RESULT ANALYSIS

Figure 6-11 shows the graph of the results obtained. Major observations of the results are as follows:

- 1. With feature selection, all three classification algorithms have shown better results in almost all three parameters.
- ZeroR has achieved better results to accurately in classifying DOS, Generic, Shellcode, Reconnaissance and worms attacks.
- IBK also performed well in classifying backdoor, shellcode and worm attacks.
- Random Forest algorithm has accurately classified most of the attacks other than reconnaissance and shellcode.
- 5. IBK has the highest attack detection rate for almost every attack.
- IBK performed well in a false attack detection rate with the lowest values for almost every attack other than backdoor and worms.
- 7. The result shows that the IBK classifier works well with an updated intrusion dataset which requires frequent updates.
- 8. Feature selection using clustering and correlation achieved better results



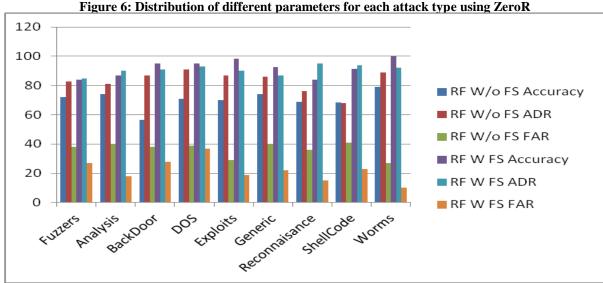


Figure 7: Distribution of different parameters for each attack type using Random Forest



Journal Website: www.ijitee.org

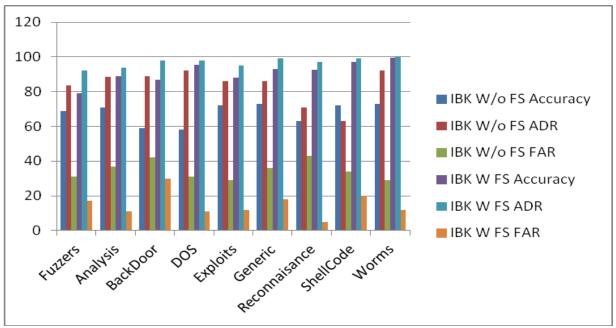
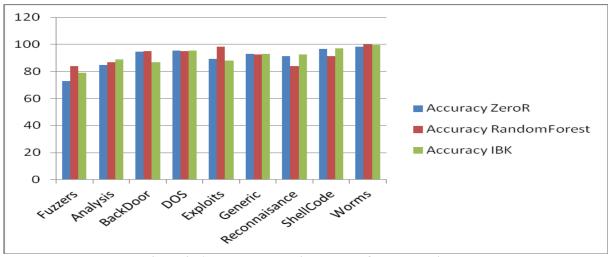


Figure 8: Distribution of different parameters for each attack type using IBK



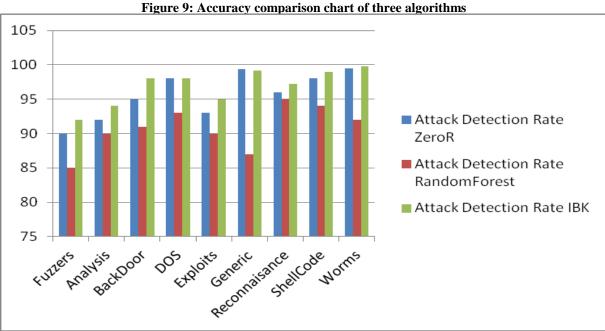


Figure 10: ADR comparison chart of three algorithms





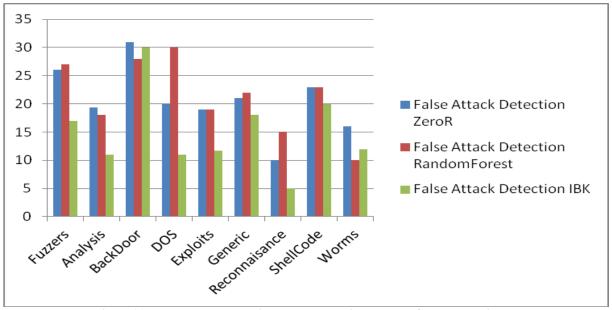


Figure 11: False attack detection rate comparison chart of three algorithms

VII. CONCLUSION

The hybrid approach is used which combines feature selection and classification algorithms. The important features were selected using clustering and attribute selection algorithms. The most significant features were: proto(5); transaction protocol which occurred in almost all attacks except Exploits and Shellcode; dur(7) Record total duration; which occurred in all expect worms, state(7) Which occurred in all except worms, sttl(10)occurs in all attacks, service(14) occurred in all except exploits and generic attacks. After feature selection, we performed classification and achieved high accuracy, attack detection rate and lower false attack detection rate In almost all type of attacks. Random forest achieved the highest accuracy in backdoor, exploits and worm attacks. IBK performed well in classifying reconnaissance attacks where accuracy is maximum and false attack rate is lowest. ZeroR failed to detect backdoor attacks and has a maximum false attack rate for backdoor attacks. Comparing all results IBK performed well after feature selection on the UNSW-NB15 dataset

REFERENCES

- Lee, W., Stolfo, S. J., and Mok, K. W. 1999. A data mining framework for building intrusion detection models. In Proceedings of IEEE Symposium on Security and Privacy. 120–132.
- Patcha, A. and Park, J.2007. An overview of anomaly detection techniques: Existing solutions and latest technological trends. Comput. Netw. 51, 12 (August 2007), 3448-3470. DOI=http://dx.doi.org/10.1016/j.comnet.2007.02.001
- Peddabachigari, S., Abraham, A., Grosan, C., and Thomas, J. 2007.
 Modeling intrusion detection system using hybrid intelligent systems.
 Journal of Network and Computer Applications. 30, 1, 114-132.
- Singh, S., Agrawal, S., Rizvi, A., and Thakur, R. S. 2011. Improved Support Vector Machine for Cyber Attack Detection. In Proceedings of the World Congress on Engineering and Computer Science. 1.
- Chitrakar, R. and Chuanhe, H. 2012. Anomaly detection using Support Vector Machine classification with k-Medoids clustering. In 2012 Third Asian Himalayas International Conference on Internet.
- Muniyandi, A., Rajeswari, R., and Rajaram, R.2012. Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm. Procedia Engineering. 30, 174-182.

- Katkar, V. D. and Kulkarni, S. V. 2013. Experiments on detection of Denial of Service attacks using ensemble of classifiers. In International Conference on Green Computing, Communication and Conservation of Energy (ICGCE), IEEE. 837-842.
- Kim, G., Lee, S., and Kim, S. 2014. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Systems with Applications. 41, 4, 1690-1700.
- Canbay, Y. and Sagiroglu, S. 2015. A Hybrid Method for Intrusion Detection. In IEEE 14th International Conference on Machine Learning and Applications (ICMLA).
- Lin, W. -C., S. Ke, S. -W., and Tsai C. -F. 2015. CANN: An intrusion detection system based on combining cluster centers and nearest neighbors. Knowledge-Based Systems. 78(Supplement C), 13-21.
- Gadal, S. M. A. M. and Mokhtar, R. A.2017. Anomaly detection approach using hybrid algorithm of data mining technique. In 2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE).
- Quinlan, J.R. 1990. Decision trees and decision-making. IEEE Transactions on Systems, Man, and Cybernetics. 20, 2, 339-346.
- 13. Govindarajan, M. and Chandrasekaran, R. 2009. Intrusion detection using k-Nearest Neighbor. In 2009 First International Conference on Advanced Computing. from: https://github.com/FransHBotes/NSLKDD-Dataset.
- Tavallaee, M., E.B., Lu, W., and Ghorbani, A. A.2009. A detailed analysis of the kdd cup 99 data set. In Proceedings of the Second IEEE International Conference on Computational Intelligence for Security and Defense Applications. 53–58.
- Chen-Mou, C., Kung, H. T., and Koan-Sin, T. 2002. Use of spectral analysis in defense against DoS attacks. In Global Telecommunications Conference. GLOBECOM '02. IEEE.
- Ahmad, I., Abdullah, A. B., and Alghamdi, A. S. 2010. Applying neural network to U2R attacks. In 2010 IEEE Symposium on Industrial Electronics and Applications.
- Bgui, S., alaimannan, E., Bagui, S., Nandi, D., & Pinto, A. (2019).
 Using machine learning techniques to identify rare cyber-attacks on the UNSW-NB15 dataset. Security and Privacy, 2(6), 1–13. https://doi.org/10.1002/spy2.91
- Gao, X., Shan, C., Hu, C., Niu, Z., & Liu, Z. (2019). An Adaptive Ensemble Machine Learning Model for Intrusion Detection. IEEE Access, 7, 82512–82521. https://doi.org/10.1109/ACCESS.2019.2923640
- Hall, M. A. (1999). Correlation-based Feature Selection for Machine Learning. April.
- Harbola, A., Harbola, J., & Vaisla, K. S. (2014). Improved intrusion detection in DDoS applying feature selection using Rank & Dcore of attributes in KDD-99 Data Set. Proceedings 2014 6th International Conference on Computational Intelligence and Communication Networks, CICN 2014. https://doi.org/10.1109/CICN.2014.179



- Mishra, P., Varadharajan, V., Tupakula, U., & Pilli, E. S. (2019). A detailed investigation and analysis of using machine learning techniques for intrusion detection. IEEE Communications Surveys and Tutorials, 21(1), 686–728. https://doi.org/10.1109/COMST.2018.2847722
- Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set).
 2015 Military Communications and Information Systems Conference, MilCIS 2015 Proceedings. https://doi.org/10.1109/MilCIS.2015.7348942
- Moustafa, N., & Slay, J. (2017). A hybrid feature selection for network intrusion detection systems: Central points. 5–13. https://doi.org/10.4225/75/57a84d4fbefbb
- Naseer, S., Saleem, Y., Khalid, S., Bashir, M. K., Han, J., Iqbal, M. M., & Han, K. (2018). Enhanced network anomaly detection based on deep neural networks. IEEE Access, 6, 48231–48246. https://doi.org/10.1109/ACCESS.2018.2863036
- Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., Gao, M., Hou, H., & Wang, C. (2018). Machine Learning and Deep Learning Methods for Cybersecurity. IEEE Access, 6, 35365–35381. https://doi.org/10.1109/ACCESS.2018.2836950

AUTHORS PROFILE



Aditya Harbola MCA, MTECH (CSE) PhD(P) engineering

- 1. An Analytical Study of Component-Based Life Cycle Models: A Survey , IEEE 12-14 Dec. 2015 DOI: 10.1109/CICN.2015.152
- 2. Green computing research challenges: a review , IJARCSSE, January 2013
- 3. Improved intrusion detection in DDoS applying feature selection using rank & score of attributes in KDD-99 data set, IEEE 14-16 Nov. 2014 DOI: 10.1109/CICN.2014.179



Dr. Priti Dimri MCA, PhD Fractals, Computer Graphics

- 1. Green computing research challenges: a review , IJARCSSE, January 2013
- Goodput Enhancement: SMR Optimization technique for improving the delivery reliability in mobile adhoc network, JIOM



Deepti Negi MCA, PhD(P)

Software engineering, Component based software engineering

1. An Analytical Study of Component-Based Life Cycle Models: A Survey , IEEE 12-14 Dec. 2015 DOI: 10.1109/CICN.2015.152

2. Green computing research challenges: a review , IJARCSSE, January 2013 , Membership: CSI, IEEE

