International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958(Online), Volume-9 Issue-5, June 2020

Host-based Intrusion Detection System (HIDS)

AmanYadav, Abhishek Srivastav, Abhinandan Tiwari, Krishna Vir Singh

Abstract: ***This paper presents the data analysis and feature extraction of KDD dataset of 1999. This is used to detect signature based and anomaly attacks on a system. The process is supported by data extraction as well as data cleaning of the above mentioned data set. The dataset consists of 42 parameters and 58 services. These parameters are further filtered to extract useful attributes. Every attack in the dataset is labeled either with "normal" or into four different attack types i.e. denial-of-service, network probe, remote-to-local or user-to-root. Using different machine learning algorithms, the work tries to compare the individual accuracy, True Positive and False positive rate of every algorithm with every other algorithm. The work focuses its attention to increase security through detection of static as well as dynamic attack.***

Keywords: Host based intrusion detection, Data cleaning, Data analysis, Machine learning, KDD cupp'99, Attack, anomaly.

I. INTRODUCTION

The paper talk about the implementation and importance of a host based intrusion detection system. This is done using refining and selection of importance features of given dataset which is achieved using data cleaning and data mining operation performed on this dataset. Some attributes contain redundant information, while some other contain false correlations; either type of these can hinder the results.
Traffic reduction is also one of the important aspects of the work. Traffic reduction can be achieved by using filters prior to network data collection. Filters ignore certain type of traffic thus reducing the traffic. Filters should be used with almost care as it can also filter or ignore some important parameters or network necessary for the determination of an attack on the host. This all data has been inculcated in the KDD dataset.All the attacks mentioned in KDD Cup '99 can be classified in any of the four attacks i.e. Denial of Service**,** REMOTE TO LOCAL, USER TO ROOT and PROBING.

A. Denial of Service

In this type of cyber-attack, attacker restricts the genuine users from accessing the services of the host which is attacked. The attack of this kind is performed by sending the huge amount of requests to the host thus making it overwhelmed and may lead to a complete crash in some cases.

Revised Manuscript Received on June 15, 2020.
* Correspondence Author
AmanYadav*, CSE Department, ABESEC Ghaziabad AKTU Lucknow , India. aman.16bcs2046@abes.ac.in.
AbhishekSrivastav, CSE Department, ABESEC Ghaziabad AKTU Lucknow , India. abhishek.16bcs2035@abes.ac.in.
AbhinandanTiwari, CSE Department, ABESEC Ghaziabad AKTU Lucknow, India. abhinandan.16bcs2051@abes.ac.in.
Krishna Vir Singh, CSE Department, ABESEC Ghaziabad AKTU Lucknow , India. krishnavir.singh@abes.ac.in.

B. Remote To Local

This kind of attack is generally performed by the attacker in order to gain illegitimate access to the targeted machine so that the attacker could steal or manipulate the data also the Attacker could inject the viruses or any kind of malicious files.

C. User To Root

In u2r attack the attacker first tries to get access to victim machine as normal user then tries to gain the access to root level after it gains the access as a root level the whole security of the targeted system is compromised and the attacker then can exploit the system at root level.

D. Probing

In this the attacker performs the scan of the entire network to find the vulnerabilities in order to find the weak points so as to gain access to the system and its files. This attack is most common because it can be performed with very little technical expertise.
This paper consists of total 22 attacks, every one of which lies in any of the four attack types e.g. pod attack is a type of DoS attack, guess_passwd attack is a R2L attack, buffer_overflow

II. MOTIVATION

Till date a lot of work has been done in the field of security but still there is a lot of scope for its improvement. No system with 100% security has been designed, there are some security flaws in every system added by all the attacks and intrusion attempt are not yet known. Tons of virus and malware are being generated everyday attacking other tons of system.
Some Recent Attacks On Indian Facilities Or On Indian Users: -

1. Agent Smith Malware in India (Aug. 2019).
2. Malware Attack on Kudankulam Nuclear Power plant (Dec. 2019).
3. WhatsApp hacking of Indian Journalist by Israeli made spyware (Nov. 2019).
4. Alert Issued by Ministry of Home against Stranghogg Bugg (Jan. 2020).
   Increasing number of cyber-attacks has increased our keen interest towards intrusion and different ways in which we can tackle them.

III. SCOPE OF HIDS

Seeing the present day scenarios intrusion attacks become a quite obvious and quite real thread for a small scale company to a large scale companies. In fact the most expanding field for now is cyber space and with its expansion the risk of attacks expands as well.

Nowadays even nations are involved in 5th generation warfare, a classic example being the latest cyber-attack carried by USA on IRAN. Thus this explains the need of HIDS in not just in companies but also for national security.

IV. RELATED WORK

Unfortunately, KDD dataset of 1999 is the best dataset available for host based intrusion detection system. Every traffic in dataset is either "normal" or any one of the four attack types. Four attacks type being denial-of-service (dos), network probe (probe), remote-to-local (r2l) and user-to-root (u2r) attacks. The data even contain anomalous behaviours of normal users acting like privileged users.

The main aim of perform a dos attack is to prevent user to access a service e.g. 'TCP syn floods'. Probe attacks like 'ipsweeps' are used to collect information about the targets. The attackers performing r2l attack try to gain admin control over machine e.g. 'dictionary attack'. In u2r attack the user with user access tries to gain privileged access. Different types of buffer overflow attacks lie in this category. Attackers can even combine different type of attack to increase their operational limits. In most of the combine attacks case the attackers go for probe→r2l→u2r pattern. Some attackers combine different attack to hide the main motive behind some other attack. For an instance many attackers perform dos attack and r2l attack, so that the user gives its attention on dos attack and r2l attack can be performed without and hindrance.

Table- I: The dataset contains 41 features along with their description

Nr	Features	
	Name	Description
1	duration	duration of connection in seconds
2	protocol_type	connection protocol (tcp, udp, icmp)
3	service	dst port mapped to service (e.g. http, ftp, ..)
4	flag	normal or error status flag of connection
5	src_bytes	number of data bytes from src to dst
6	dst_bytes	bytes from dst to src
7	land	1 if connection is from/to the same host/port; else 0
8	wrong_fragment	number of 'wrong' fragments (values 0,1,3)
9	urgent	number of urgent packets
10	hot	number of 'hot' indicators (bro-ids feature)
11	num_failed_logins	number of failed login attempts
12	logged_in	1 if successfully logged in; else 0
13	num_compromised	number of 'compromised' conditions
14	root_shell	1 if root shell is obtained; else 0
15	su_attempted	1 if 'su root' command attempted; else 0
16	num_root	number of 'root' accesses
17	num_file_creations	number of file creation operations
18	num_shells	number of shell prompts
19	num_access_files	number of operations on access control files
20	num_outbound_cmds	number of outbound commands in an ftp session
21	is_hot_login	1 if login belongs to 'hot' list (e.g. root, adm); else 0
22	is_guest_login	1 if login is 'guest' login (e.g. guest, anonymous); else 0
23	count	number of connections to same host as current connection in past two seconds
24	srv_count	number of connections to same service as current connection in past two seconds
25	serror_rate	% of connections that have 'SYN' errors
26	srv_serror_rate	% of connections that have 'SYN' errors
27	rerror_rate	% of connections that have 'REJ' errors
28	srv_rerror_rate	% of connections that have 'REJ' errors
29	same_srv_rate	% of connections to the same service
30	diff_srv_rate	% of connections to different services
31	srv_diff_host_rate	% of connections to different hosts
32	dst_host_count	count of connections having same dst host
33	dst_host_srv_count	count of connections having same dst host and using same service
34	dst_host_same_srv_rate	% of connections having same dst port and using same service
35	dst_host_diff_srv_rate	% of different services on current host
36	dst_host_same_src_port_rate	% of connections to current host having same src port
37	dst_host_srv_diff_host_rate	% of connections to same service coming from diff. hosts
38	dst_host_serror_rate	% of connections to current host that have an S0 error
39	dst_host_srv_serror_rate	% of connections to current host and specified service that have an S0 error
40	dst_host_rerror_rate	% of connections to current host that have an RST error
41	dst_host_srv_rerror_rate	% of connections to the current host and specified service that have an RST error
42	connection_type	

There are basically three methodology or neural networks that are used in implementation of host based intrusion detection system: -

A. Recurrent Neural Network

It is a form of network with backward connection. The output layer in the network is fed back into either that layer or the previous layer in the network.

B. Convolutional Neural Network

It is primarily used in image processing. It is used to detect patters. This network helps in the identification of attack patterns or change in a normal pattern.

C. Sequence anomaly detection using language modeling

In this system call is represented as integers (1 to 340). We can estimate probability of sequence occurring using probability distribution.

This network helps in estimation of attack based on the sys logs and pattern analysis.The KDD Cup '99 gives us idea about some errors as well. The errors include syn error, rej error, S0 errorandrst error.

Table- II: Different attacks that are classified under the four attack categories

Name of the attack	Type	Mechanism	Effect of the attack
back	DoS	Abuse/Bug	Slows down server response
land	DoS	Bug	Slows down server response
neptune	DoS	Abuse	Slows down server response
smurf	DoS	Abuse	Slows down the network
pod	DoS	Abuse	Slows down server response
teardrop	DoS	Bug	Reboots the machine
loadmodule	U2R	Poor environment sanitation	Gains root shell
buffer_overflow	U2R	Abuse	Gains root shell
rootkit	U2R	Abuse	Gains root shell
perl	U2R	Poor environment sanitation	Gains root shell
phf	R2L	Bug	Executes commands as root
guess_passwd	R2L	Login misconfiguration	Gains user access
warezmaster	R2L	Abuse	Gains user access
imap	R2L	Bug	Gains root access
multihop	R2L	Abuse	Gains root access
ftp_write	R2L	Misconfiguration	Gains user access
spy	R2L	Abuse	Gains user access
warezclient	R2L	Abuse	Gains user access
satan	Probe	Abuse of feature	Looks for known vulnerabilities
nmap	Probe	Abuse of feature	Identifies active ports on a machine
portsweep	Probe	Abuse of feature	Identifies active ports on a machine
ipsweep	Probe	Abuse of feature	Identifies active machines

Massachusetts Institute of Technology Lincoln labs gave KDD dataset, which is found to be very helpful for framing any machine learning model we in our project have used four machine learning algorithms for classifying various attacks into 4 broad categories. Each of the
machine learning algorithms has its own accuracy for each type of categories. The statistical information present in the dataset contains sufficient amount of instances for each type of attacks.It makes very easy to divide the training and testing dataset.

V. IMPLEMENTATION AND RESULT

A. Machine Learning Approach

Approach of machine learning is very simple in a machine learning we try to design a system that is able to complete the task by closely studying the training dataset The formed predictive model then based upon the findings during training dataset classifies the test data into various labels. Accuracy and reliability of the predictive model increases with the increase in the amount of data present and this demand is fulfilled by the KDD dataset. Thus it makes very obvious to follow the machine learning approach.

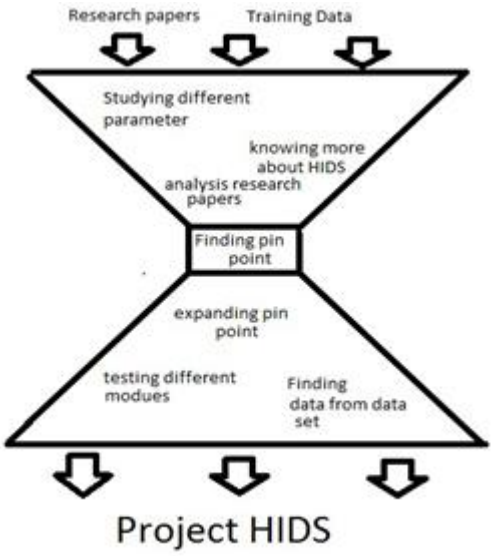


Fig.1. Anomaly detection flowchart using machine learning algorithms

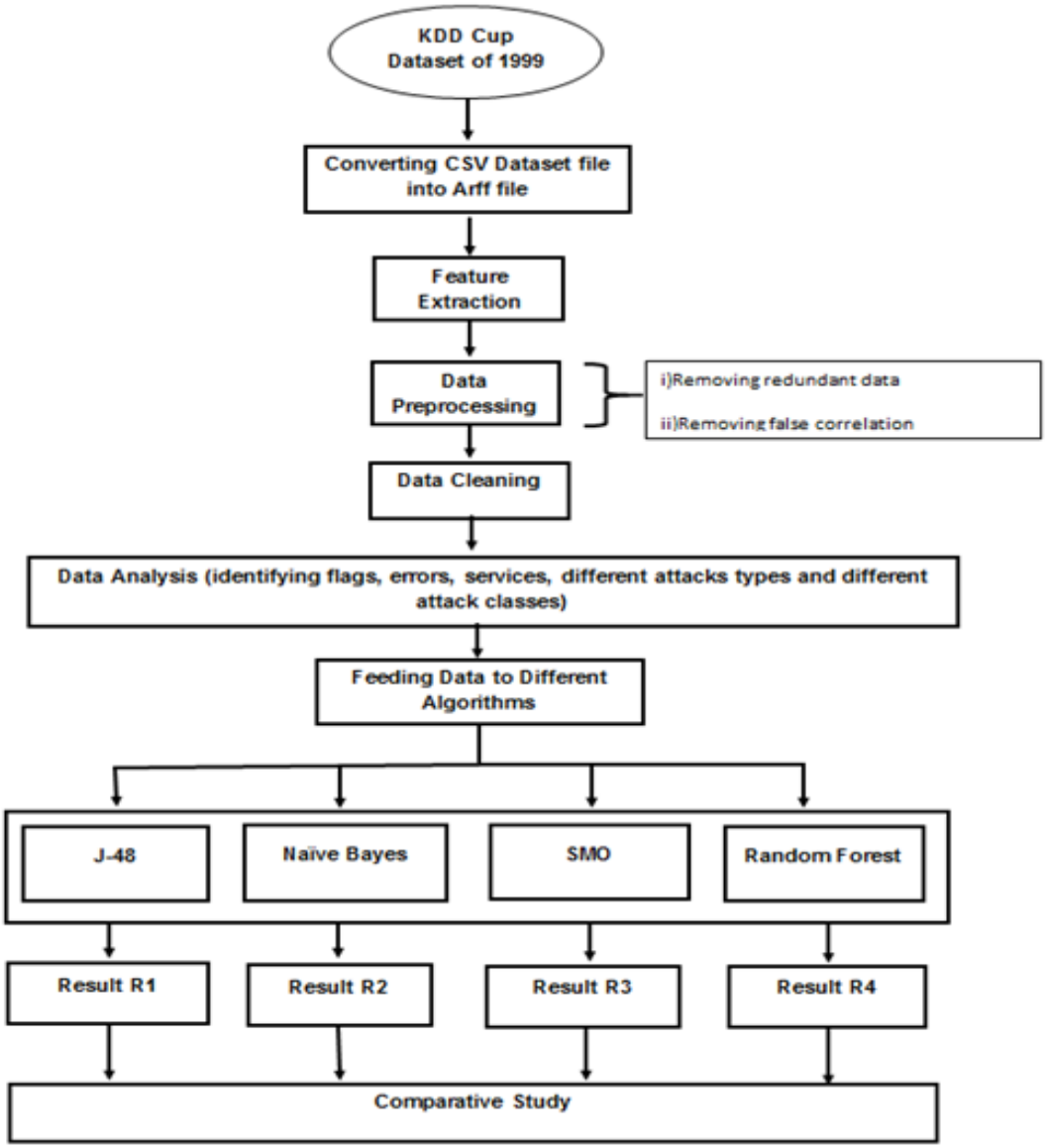


Fig.2. the above flowchart clearly illustrates the steps required to accomplish the task.

B. Comparative study (Accuracy of different Algorithms)

The paper involves use of four algorithms i.e. Naïve Bayes, J-48, Random forest and SOM algorithm. Each algorithm generates result output with different accuracy and these accuracies are compared in order understand
the efficiency of every algorithm with KDD Cup '99 datasets. The output generated contains the TP rate and FP rate. TP rate stands for true positive rate and it measures the actual positives that are correctly identified. The other name for TP rate is "sensitivity" .FP rate stands for false positive rate and it is given by following expression: -

$$(FP / N) = FP/(FP+TN) \quad (1)$$

$$N = FP + TN \quad (2)$$

Where FP is the number of false positives, TN is the number of true negatives and N is the total number of negatives.

Table-III: TP rate and FP rate of algorithms for every attack level

Class	NaiveBayes		J48		Random Forest		SOM	
	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate	TP Rate	FP Rate
Normal	71.2	0.3	99.9	1	100	0.8	99.9	2.6
buffer_overflow	80	0.9	70		86.7	0	53.3	0
loadmodule	33.3	0.1	0	0	22.2	0	0	0
perl	33.3	0	0	0	66.7	0	33.3	0
neptune	98.6	0.1	99.9	0	100	0	100	0
smurf	99	6.6	100	0	100	0	91.1	0
guess_passwd	94.3	0.1	94.3	0	96.2	0	96.2	0
pod	98.9	0.6	98.9	0	98.9	0	98.9	0
teardrop	100	0	100	0	100	0	99.4	0
portsweep	44.9	0.27	97.7	0	96.7	0	97.2	0
ipsweep	97.4	5	97.9	0	96.7	0	97.4	0.1
land	94.4	0	83.3	0	94.4	0	100	0
ftp_write	50	0.2	0	0	50	0	12.5	0
back	98.4	5.7	99.2	0	100	0	98.8	0
imap	91.7	0	25	0	91.7	0	83.3	0
satan	84	1	96.2	0	95.1	0	89.2	0
phf	100	0.7	100	0	100	0	0	0
nmap	31.7	0.7	93.1	0	96.6	0	32.4	0
multihop	28.6	0	28.6	0	28.6	0	0	0
warezmaster	80	0	75	0	85	0	75	0
warezclient	49.8	2.8	98.6	0	98.8	0	91.6	0
spy	100	0	0	0	0	0	0	0
rootkit	30	0.3	0	0.9	10	0.7	0	0

The classifier values are different for all the four algorithms i.e. correctly classified instances, incorrectly classified instances, mean absolute error, root mean squared error, relative absolute error, root relative squared error and total numbers of instances. After applying various classifying algorithms (Random Forest, SMO, Naïve Bayes, J48), it has been found that the Random Forest algorithm has been successful in detecting various attacks in KDD Dataset and classifying instances of KDD dataset correctly.

Table-IV: Classifier values for different algorithms

Classifier	Naïve Bayes		J48		Random Forest		SOM	
Correctly Classified Instances	89739	99.80%	65655	73.0165 %	89794	99.8621 %	89455	99.4851 %
Incorrectly Classified Instances	179	0.20%	24263	26.9835 %	124	0.1379 %	463	0.5149 %
Mean absolute error	0.0002		0.0232		0.0003		0.0794	
Root mean squared error	0.0125		0.1469		0.0099		0.1961	
Relative absolute error	1.42%		143.81%		1.81%		492.51%	
Root relative squared error	13.90%		163.69%		11.04%		218.59%	
Total Number of Instances	89918		89918		89918		89918	

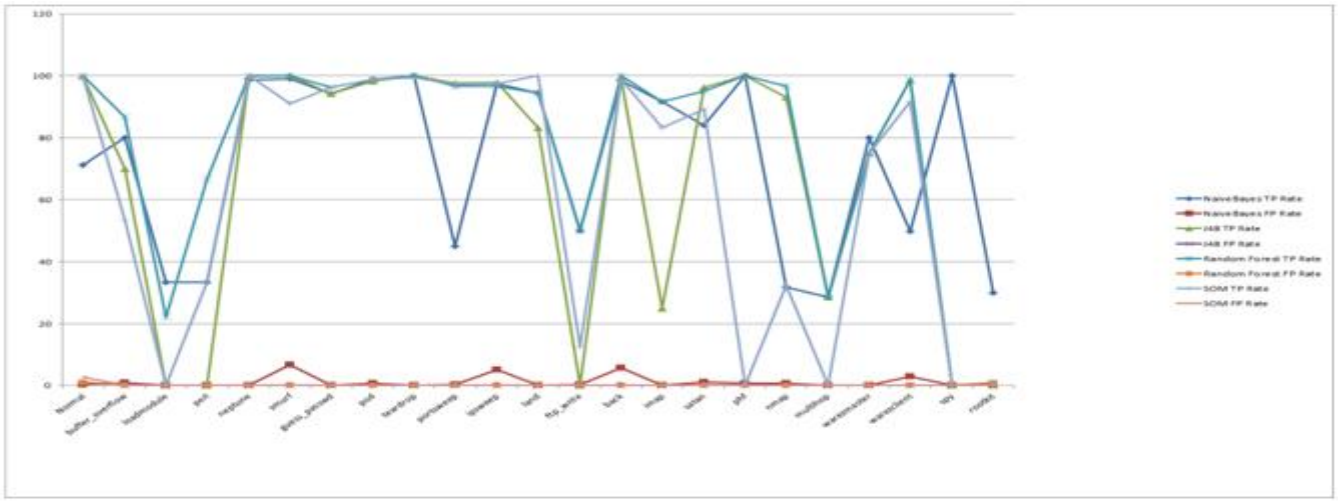
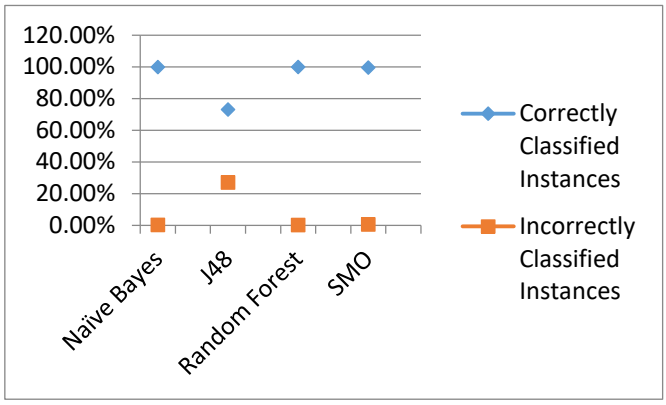
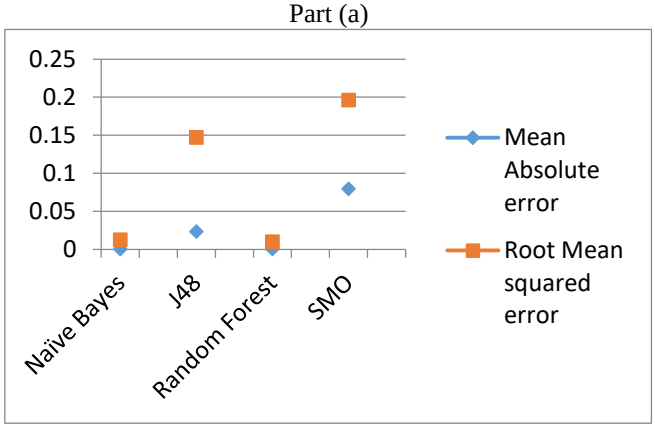


Fig. 3. The below graph represent FP and TP rate for different algorithms on different attacks



Part (a)



Part (b)

The classifier values and errors for different algorithms represented in form of graph

VI. CONCLUSION AND FUTURE SCOPE

Because of the acute demand of an effectual HIDS in the security related to the network, many researchers are trying very hard and are working for better approach that can lead to better results. In this research paper depicts the usefulness of the KDD dataset for testing various classifiers. Study focuses on preprocessing of KDD Dataset and a lot of work has been done to remove all sorts of factors that can lead to bias results. Originally KDD has 42 attributes and after studying all of them in detail and finding which of them are required for HIDS we have reduced them to 31 parameters. Now talking about the future scope there is a lot work which is need to be done to increase the efficiency of the classifying algorithms we can tune the algorithms we can alter the % of training and testing data. We need to increase the efficiency of the HIDS.

REFERENCES

1. Staudemeyer, Ralf C., and Christian W. Omlin. "Extracting salient features for network intrusion detection using machine learning methods." South African computer journal 52.1 (2014): 82-96.
2. Akbar, Shaik, K. NageswaraRao, and J. A. Chandulal. "Intrusion detection system methodologies based on data analysis." International Journal of Computer Applications 5.2 (2010): 10-20.
3. Win, MyaThidarMyo, and KyawThetKhaing. "Detection and Classification of Attacks in Unauthorized Accesses." International Conference on Advances in Engineering and Technology (ICAET'2014). 2014.
4. Sabhnani, Maheshkumar, and GürselSerpen. "KDD Feature Set Complaint Heuristic Rules for R2L Attack Detection." Security and Management. 2003.
5. Chawla A, Lee B, Fallon S, Jacob P. Host based intrusion detection system with combined CNN/RNN model. InJoint European Conference on Machine Learning and Knowledge Discovery in Databases 2018 Sep 10 (pp. 149-158). Springer, Cham.
6. Bace RG. Intrusion detection. Sams Publishing; 2000.
7. Olusola, A.A., Oladele, A.S. and Abosede, D.O., 2010, October. Analysis of KDD'99 intrusion detection dataset for selection of relevance features. In Proceedings of the world congress on engineering and computer science (Vol. 1, pp. 20-22). WCECS.
8. Aggarwal P, Sharma SK. Analysis of KDD dataset attributes-class wise for intrusion detection. Procedia Computer Science. 2015 Jan 1;57:842-51.
9. Lee, J.H., Lee, J.H., Sohn, S.G., Ryu, J.H. and Chung, T.M., 2008, February. Effective value of decision tree with KDD 99 intrusion detection datasets for intrusion detection system. In 2008 10th International Conference on Advanced Communication Technology (Vol. 2, pp. 1170-1175). IEEE.
10. Chae HS, Jo BO, Choi SH, Park TK. Feature selection for intrusion detection using NSL-KDD. Recent advances in computer science. 2013 Nov:184-7.
11. Lakhina, Shilpa, Sini Joseph, and BhupendraVerma. "Feature reduction using principal component analysis for effective anomaly–based intrusion detection on NSL-KDD." (2010).

AUTHOR PROFILE

Mr. Aman Yadav is pursuing his B.Tech in Computer Science and Engineering from ABES Engineering College, Ghaziabad which is affiliated to AKTU. He has completed his 10th and 12th from Mercy Memorial School, Kanpur affiliated to ICSE. He has worked in java based technologies and J2EE technologies. He is currently working in the domain of cyber security and computer networking. He is very eager to learn new technologies because it further expands the opportunities upon which he can work upon and solve the problems. He believes that different technologies can be brought together in order to make a product reliable.

Mr. Abhishek Srivastav from Gorakhpur city of Uttar Pradesh. My family which includes his mother, father and sister have always provide their support in his journey as a student and as an individual.His father is a government officer and his mother is a home maker. In the way to his graduation from ABES Engineering college in Computer science, he has completed his high school and intermediate from Little Flower School (Salempur) and Sunbeam Bhagwanpur(Varanasi) with 83% and 93% respectively.He had a great enthusiasm towards computers that helped him in choosing his path for graduation. Computer and cyber security had always been in Abhshek's mind since he first started using computers. Moving up in his life he had looked forward to expand his knowledge and his work in the field of computers and their security.

Mr. Abhinandan Tiwari is currently pursuing his B.tech in Computer Science and Engineering from ABES Engineering College which is affiliated to AKTU. He has completed his 10th and 12th from Bright Way College Lucknow with 8.4 CGPA and 83.4%. He worked on JAVA based technologies and J2EE technologies .He is currently working is the domain of networking and cyber security. Computer and cyber security had always been in my mind since he first started using computers. Moving up in my life i look forward to expand my knowledge and my work in the field of computer networking and cyber security.

Mr. Krishna Vir Singh is Currently Associated with ABES Engineering College, Ghaziabad as Assistant Professor. He has More Than 11 Years of Experience in Education, Training and Software Development Industry. He is CISCO Certified Instructor for CCNA (Routing & Switching). CCNA(Cyber Ops), IoT Fundamentals : Connecting Things, IoT Fundamentals : Big Data Analytics, IoT Security and he is also a ICSI Certified Network Security Specialist. His Research Interests includes Genetic Algorithms, Software Testing, IoT, Network Security & Cyber Security. He has also filed patent in the area related to IoT.