Improved Network Intrusion Detection Systems (NIDS) using Data Mining Techniques

Arun Pandey[1], Ayush Kumar Agrawal[2]

[1]Research Scholar, Dr. C. V. Raman University, Kota, Bilaspur (C.G), India

[2]Assistant Prof. and HoD Dept. of IT and CS, Dr. C. V. Raman University, Kargi Road Kota, Bilaspur (C.G), India

E-mail: [1]arun.pandey151989@gmail.com ,[2]ayushagrwal369@gmail.com

Abstract: **-**The subject of Intrusion Detection System (IDS) is a very interesting research topic actively pursued by many investigators. The goal of intrusion detection is to monitor network assets and to detect anomalous behaviour and misuse. Intrusion Detection Systems aim to identify attacks with a high detection rate and a low false alarm rate. Intrusion Detection Systems (IDS) can monitor users, applications, networks, or combinations of the three, in order to detect well-known and unknown attacks. In this research work, I read many papers in which I found that some papers used supervised machine learning method. In this method, SVM algorithm was used and kernel function was also used. Using this concept, when intrusion detection system was built, its accuracy was between 70% to 81%. Apart from this, when I run generic algorithm and another model based on signature, its performance was between 85% to 91%. Similarly, when I used generic algorithm and another model based on anomaly, its performance was the best. Apart from this, I run some other models in which unsupervised method was used. In this, FCM algorithm was used which follows DBSCAN algorithm.it reduce the false positive rate. In my research paper, I have used a hybrid method in which I have created an intrusion detection model using SVM, FCM and DBSCAN algorithm which not only reduces the false positive rate but also improves network security.

Keywords: Data Mining, Intrusion Detection, Classification, False Positive, Anomaly based algorithm, Machine Learning, Deep Learning, NSL-KDD Data set.

1. INTRODUCTION TO INTRUSION DETECTION SYSTEM

The Intrusion Detection Systems inspect the incoming and outgoing traffic of a system or a network for doubtful action and alert the system administrator of all possible intrusions. IDS raises alarms only when the user action significantly deviates from baseline behaviour or match with signature of a known malicious content allowing the user to block such applications.

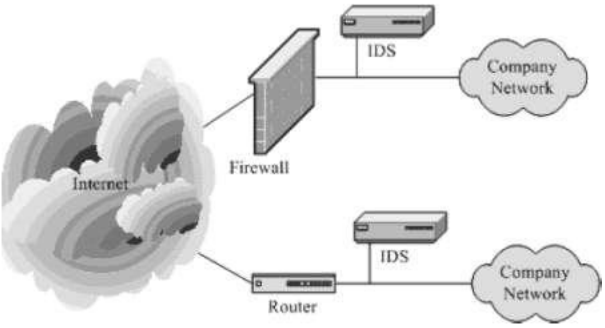


Figure 1.1

1.1 Types of Intrusion Detection Systems:

Intrusion Detection Systems are categorized based on the data being monitored and methodology adopted. Based on the protection provided, they are classified as a Network-based Intrusion Detection System and Host-based Intrusion Detection System.

1.1.1 Network-based Intrusion Detection System (NIDS)
1.1.2 Host-based Intrusion Detection System (HIDS)
1.1.2.1 Signature-based detection
1.1.2.2 Anomaly-based detection.
1.1.2.3

1.1.1Network-based Intrusion Detection System (NIDS)

The importance of Network Intrusion Detection System (NIDS) is a system that tries to identify misbehaviour activity like as denial of service attacks, scanning of ports or even trying to crack into computers by Network Security Monitoring (NSM) of network traffic. NIDS scans all the incoming packets and tries to find unauthorized patterns called as signatures or rules.

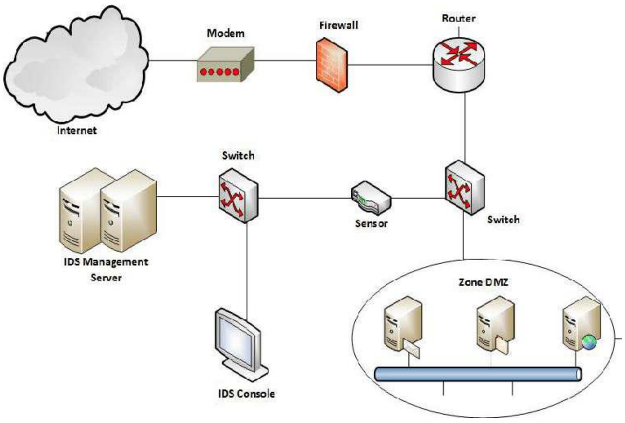


Figure 1. Network-based intrusion detection system

The NIDS then scans any traffic that is flowing over that segment of the network, as shown in Figure 1.3. The NIDS function reflects the same way as high-end antivirus applications, and it is comparing each transmitted packet against the signature or pattern file method. The IDS functions in a concrete way to increase packet throughput as inspecting every packet can slow traffic considerably. Further, An IDS uses the firewall methodology while checking the packet by letting through the packets that are not harmful to the system. Preprocessing filters carried out this task for IDS.

1.1.2. Host-based Intrusion Detection System (HIDS)

Much as a NIDS will dynamically verify the network packets, HIDS are run on individual hosts or machine on the network. A HIDS tracks the inbound and outbound packets from the device only and alerts the user or administrator of abnormal activity is detected.

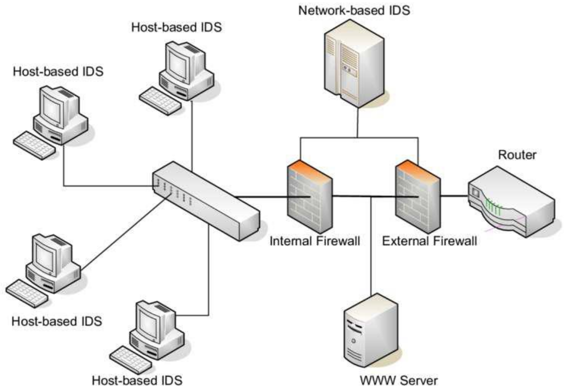


Figure 2. Host-based Intrusion Detection System (HIDS)

1.2. KDD CUP 1999 Data set

The KDD Cup 1999 dataset (often called KDD99) is one of the most famous datasets in machine learning, particularly for network intrusion detection tasks.

Categories	Training Data	Testing Data
Normal	97278	60593
DOS	391458	223298
R2L	1126	5993
U2R	52	39
Probing	4107	2377
Others	0	18729
Total	494021	311029

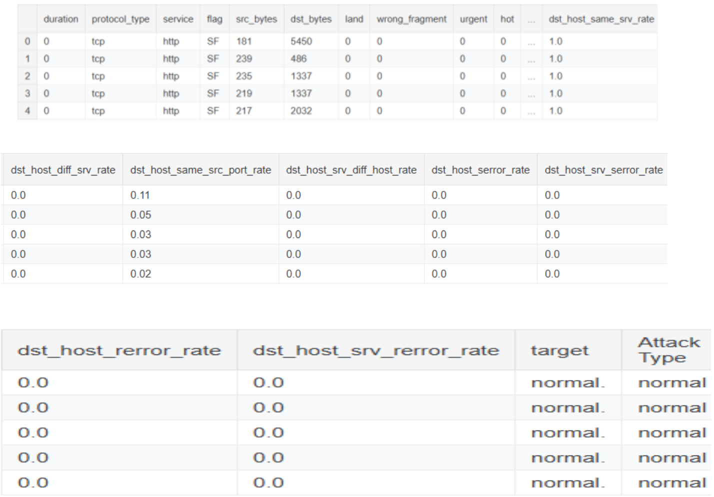
Table 1. The KDD CUP 1999 Data Set

Attack types fall into 4 broad categories:

- DoS (Denial of Service) — e.g., SYN flood
- R2L (Remote to Local) — unauthorized access from a remote machine
- U2R (User to Root) — unauthorized access to local superuser privileges
- Probe — surveillance and probing (e.g., port scanning)

Attacks	**Types**
Denial Of Service	Back, Land, Neptune, Teardrop, Smurf, Pod
Probe	Ipsweep, Portsweep, Satan, Nmap
User To Root	Buffer _Overflow, Load Module, Perl, Rootkit
Remote to Local	Multihop, Warezclient, Warezmaster, Ftpwrite, Imap, Phf, Gues_Passwd,Spy
Other	Normal

Table 2. The KDD CUP 1999 Data Set (Attacks)

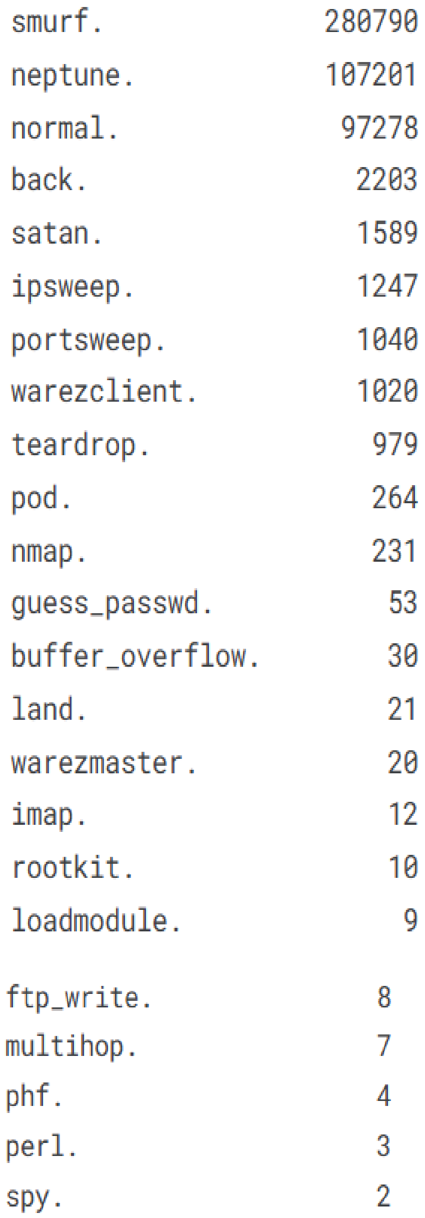


	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot	...	dst_host_same_srv_rate
0	0	tcp	http	SF	181	5450	0	0	0	0	...	1.0
1	0	tcp	http	SF	239	486	0	0	0	0	...	1.0
2	0	tcp	http	SF	235	1337	0	0	0	0	...	1.0
3	0	tcp	http	SF	219	1337	0	0	0	0	...	1.0
4	0	tcp	http	SF	217	2032	0	0	0	0	...	1.0

dst_host_diff_srv_rate	dst_host_same_src_port_rate	dst_host_srv_diff_host_rate	dst_host_serror_rate	dst_host_srv_serror_rate
0.0	0.11	0.0	0.0	0.0
0.0	0.05	0.0	0.0	0.0
0.0	0.03	0.0	0.0	0.0
0.0	0.03	0.0	0.0	0.0
0.0	0.02	0.0	0.0	0.0

dst_host_rerror_rate	dst_host_srv_rerror_rate	target	Attack Type
0.0	0.0	normal.	normal
0.0	0.0	normal.	normal
0.0	0.0	normal.	normal
0.0	0.0	normal.	normal
0.0	0.0	normal.	normal

Fig 3. The KDD CUP 1999 Data Set (Reading Data Set)



```
smurf.             280790
neptune.           107201
normal.             97278
back.                2203
satan.               1589
ipsweep.             1247
portsweep.           1040
warezclient.         1020
teardrop.             979
pod.                  264
nmap.                 231
guess_passwd.          53
buffer_overflow.       30
land.                  21
warezmaster.           20
imap.                  12
rootkit.               10
loadmodule.             9

ftp_write.            8
multihop.             7
phf.                  4
perl.                 3
spy.                  2
```

Fig 4. The KDD CUP 1999 Data Set (Value Count)

2. Literature Survey

S. N.	DATA SET	TECHNIQUES	ALGORITHM	RESULT
1	NSL-KDD	Signature Based	Genetic Algorithm	81.61%,65.40%,61.30%,55.43%
2	1998 DARPA BSM	Anomaly Based kNN Classifier	Genetic Algorithm	81.8%,81.8%,63.6%
3	1998 DARPA BMS	K-Nearest Classifier	Genetic Algorithm	75%

4	NSL-KDD	GA-BFSS and Logistic Regression	Genetic Algorithm	74.94
5	NSL-KDD	Neural Network Back Propagation	K-Means	Misuse and Anomaly Instruction
6	KDD 999	Neural Network Back Propagation	K-Means	Less Complexity, Low Processing Time, Good Generalization ability
7	DARPA	Artificial Neural Network and Neural Network Back Propagation	K-Means	Unknown Malicious User and 76% Accuracy
8	DARPA	Artificial Neural Network and Multi class classification, multi-layer perceptron with error back propagation	K-Means	91% and 87%
9	DARPA	Multi-layer perceptron without error back propagation	K-Means	91%
10	NSL-KDD	Artificial Neural Network	K-Means	81.2%
11	KDD cup99	Support Vector Machine and Artificial Neural Network	K-Means	92.20%,90.60%
12	KDD 99	Multi-Layer Perceptron with back propagation	Apriori	94%
13	KDD 99	Artificial Neural Network	K-Means	High Accuracy 97.66%
14	NSL-KDD	Artificial Neural Network	K-Means Cluster	High Detection Rate and Low False Alarm
15	NSL-KDD	Anomaly	K-Means Clustering	High Accuracy 98.67%

3. Clustering Algorithm

3.1. Introduction To Data Mining

Data mining is the process of bringing out valid, authentic, and actionable information from large databases. It is a logical process designed to explore data in search of consistent patterns and well-ordered relationships between variables, and then to confirm the findings by appealing the detected patterns against the new subsets of data. Prediction is the supreme goal of data mining - and predictive data mining is one of the primary types of data mining, and one that has connected with most direct business applications. The KDD cup data set process consists of many steps. They are,
Data Cleaning - removing noise and inconsistent data

Data Integration - multiple data sources may be combined.

Data Selection - data relevant to the analysis task are retrieved from the

database.

Data Transformation - data are consolidated into forms appropriate for mining

by performing or aggregation operations.

Data Mining - Intelligent methods are applied in order to extract data patterns.

Pattern Evaluation - Identify the truly interesting patterns representing

knowledge based on certain measures.

Knowledge Representation - Visualization of the mined knowledge to the user.

3.2. Need For Support Vector Machine

A Support Vector Machine (SVM) algorithm belongs to the supervised learning methods. To examine the data and recognize patterns, which used for classification and regression analysis process. The standard SVM algorithm considers a set of input data and predicts, for every input. Support Vector Machine models are a close cousin to classical Multi- Layer Perceptron Neural Networks (MLPNN). Using a kernel function, SVM's are a substitute training method for polynomial, radial basis function, and Multi-Layer Perceptron (MLP) classifiers. In which the network weights are found by resolving quadratic programming problem with linear constraints, rather than by solving a non-convex, the unconstrained minimization problem as in standard neural network training.

The main goal of SVM is to reduce the training dataset by classifying the dataset into the positive and negative kernel using radial basis function. Classifying data is a general task in machine learning. Suppose few given data points each associated with one of two classes, and the target is to finalize which class a new data point is generated, and it's shown in Figure 3.1

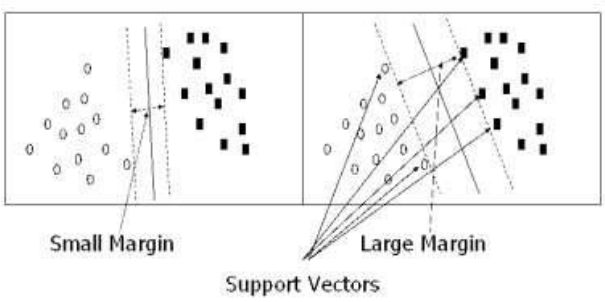


Figure 5. Support Vector Machine

Incremental SVM algorithm to deal with network intrusion detection, called RS-ISVM. Firstly, to reduce the noise generated by feature differences, we suggest an improved incremental kernel function U-RBF, which is based on kernel function Gauss kernel function (RBF). Secondly, to ease the oscillation phenomenon that often occurs in the learning process of simple incremental SVM develops a reserved set strategy to retain those non-support vectors that are likely to be the support vectors. Concentric circle method is employed to select the samples to construct the reserved set.

3.3. Clustering Technique

Clustering is a process which splitting a given data set into equivalent groups based on given features such that similar objects are included in a group whereas non identical objects are in different groups. It is the superior unsupervised learning problem. It deals with the identical structure in a collection of unlabelled data. For better understanding, please refer to Figure 4.2

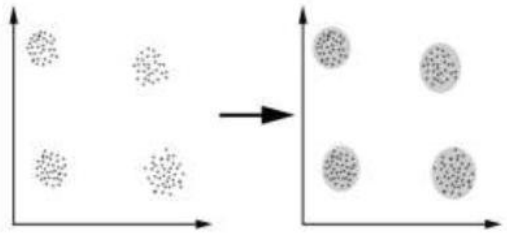


Figure 6. Cluster formed the set of unlabelled data

4. Proposed System for Network Intrusion Detection and Prevention System

The proposed system is an intelligent Anomaly-Based Network Intrusion Detection and Prevention System (NIDPS) designed to detect, classify, and block malicious activities in real time. It combines machine learning techniques with adaptive models to achieve high detection rates while minimizing false positives.

It is a security system that monitors network traffic in real time, detects abnormal (unexpected) patterns that could indicate a cyberattack, and prevents harm by taking action — like blocking traffic or alerting administrators.

Instead of looking for known attack signatures (like signature-based systems do), anomaly-based systems learn what normal behavior looks like and detect deviations from that normal pattern.

An Anomaly-based Network Intrusion Detection and Prevention System architecture

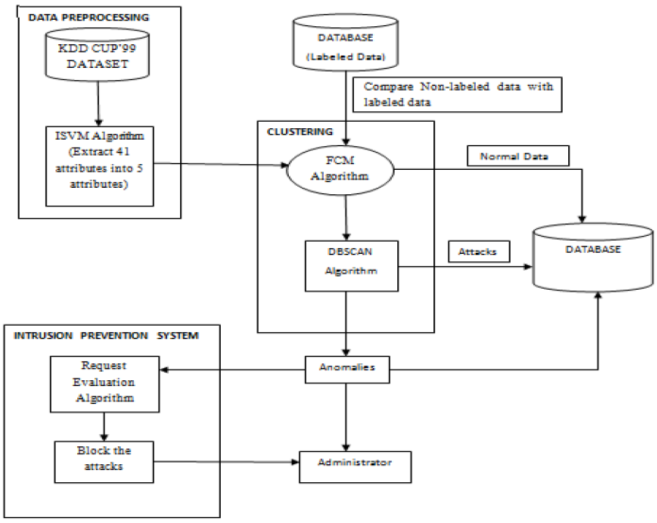


Figure 7. Proposed System for ANIDPS

4.2 Implementation of Proposed Scheme

In this work, the Support Vector Machine is used to pre-process the data using kernel function. It is used to classify the data set into two types of datasets named as positive kernel and negative kernel. After pre-processing, the FCM algorithm is applied to that data set, which is used to differentiate the abnormal data and the normal data. FCM algorithm is followed by the DBSCAN algorithm, which is used to compare the data in the normal profile with unlabelled data and form the clusters on the basis of types of attacks. The data which are not included in the any of the above clusters is called as anomalies. Detected threads or attacks are given as an input to the IPS. If the given request is detected as an attack or anomaly, then it blocked by the administrator.

4.3. **Data Pre-Processing**

The SVM is used to calculate the kernel function of every input data using RBF kernel function, it leads to reduces the number of input data given to the next process. unlabelled data given as an input, which is taken from the KDD

$K(x,y)=exp(-((x-y)2)/2\sigma2)$

In this method, the formula for kernel function as

$K(x,y)=exp(-((x-m)/s)-((y-m)/s))2/2\sigma2)$

The radial function of every input data is computed by calculating the mean and median of any two attributes. The mean or medians are calculated by taking source bytes and destination bytes as inputs. Source bytes, destination bytes, and count are named as x, y and s respectively. CUP 1999 dataset. Usually, the kernel function uses the formula

The mean value is calculated by, n

$Mj=1/n$

where, m = (m1, m2, . . . mj, . . .,mk) and s = (s1, s2, . . .sj, . . .,sk) are the mean value and the mean square deviation of attributes, k is the dimension of sample vectors.

The square deviation can be calculated by,

$Sj=(1/n-1)$

The kernel function yields two results positive kernel value and negative kernel value. Positive kernel values, which also called as reduced unlabelled data, will be taken for further computation and negative kernel values will be ignored, which in turn reduce the sparsity.

Algorithm for Implementing ISVM

```
Input: KDD CUP 1999 Dataset
Output: Sample unlabeled dataset
Procedure:
Start
For each unlabeled dataset
Do
Find mean for unlabeled dataset M;
Find mean square deviation for unlabeled dataset S;
Find kernel function k(u,v);
If(k(u,v)>0)
Store the dataset in the database
End if
Else
Reject the dataset
End else
end for
end
```

4.4. **FCM Algorithm**

Positive kernel value obtained from the pre-processing step is given as input to the FCM algorithm. This algorithm compares the input data (unlabelled data) with the data in the normal profile. Usually, the normal profile consists of data which has been clustered based on the

Euclidian distance. Unlabelled data (input data) compared with the normal profile has been placed in the corresponding clusters. Finally, the data which has not been placed in any of the above clusters will be called abnormal data.

Algorithm for implementing FCM

```
Input: Attacks
Output: Specification of attacks or anomalies
Procedure:
Start
For each unlabeled dataset
Extract the src_bytes and dest_bytes
Do
Compare the src_bytes and dest_bytes with the normal profile
If( unlabeled dataset==normal profile)
Add the dataset to the normal profile
end if
else
Find_attacks;
end else
end for
end
```

4.5. DBSCAN Algorithm

FCM algorithm applied to the pre-processed data will result in finding the abnormal data. The abnormal data which is also called as unlabelled data is given as input to this algorithm. This data again compared with the data in the normal profile, which contains labelled data. The abnormal data has been placed in the clusters available in the normal profile based on source bytes and destination bytes. The resultant data which has not been placed in any of the above clusters is called as anomalies. This algorithm leads to reduce the false positive rate.

Algorithm for implementing DBSCAN

```
Input: Attacks
Output: Specification of attacks or anomalies
Procedure:
Start
Store the labeled attacks in the database
for each attacks
Do
Compare the attacks with the labeled attacks
If ( src_bytes= " " && dest_bytes= " ")
Display the attack
End if
Else
Display it as anomaly
End else
Find the attacks and add them to the corresponding cluster
End for
End
```

4.6. Real-Time Implementation

Using Smart sniff tool, monitored the network traffic and captured the packets. From the real-time data, extract the Loc_port, Remote_port, and src_bytes. Then Calculate the mean and variance for loc_port, Remote_port, src_bytes. Cluster the data according to the mean and variance of Loc_port and Remote_port. For getting more efficient output, again clustered the data based on the distance for Local_port, Remote_port, and src_bytes.

Real Time Implementation Algorithm

```
Input: Real-time Unlabeled data
Output: An indication of attacks or normal
Procedure:
Start
For each real-time data
Do
Extract the Local_port, Remote_port and src_bytes
Calculate the mean for Local_port, Remote_port and Src_bytes;
Calculate the variance for Local_port, Remote_port and
Src_bytes;
Cluster the data according to mean and variance
```

```
If (Local_port<="0.0025" && Remote_port<= "0.0026")
Then the Given data is normal
End if
Else
Attack
End else
end for
end
```

4.7. Performance Evaluation

The implemented Anomaly-based Network Intrusion Detection system (ANIDS) that integrates the SVM and Hybrid Clustering Algorithms and the performance is evaluated and compared with the results of FCM Algorithm. The result shows the performance is better than the existing system. The detection rate of an Anomaly-based Network Intrusion Detection System that uses the KDD Cup 1999 Dataset is comparatively higher than the existing system. The detection rate of ANIDS

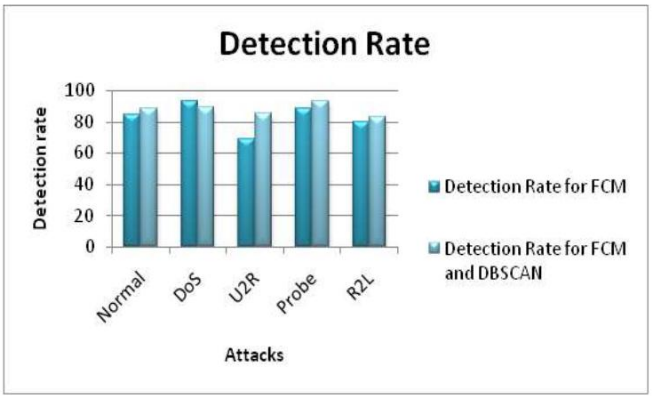


Figure 8. Detection Rate of anomaly -Based NIDS

A false-positive rate for FCM is individually calculated and the performance of false positive rate calculated with the combination of FCM and DBSCAN also estimated against different kind of attacks like normal, DoS, etc and the results are compared to the existing system. It is shown below

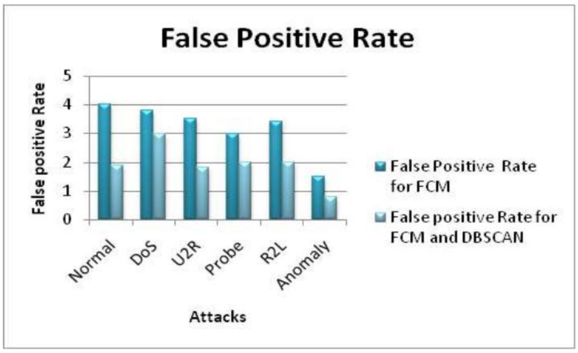


Figure 9. False Positive Rate of anomaly -Based NIDS

The performance is better than the existing system. Using the KDD Cup 1999 Data set with five attributes for analysis is relatively better than handling with nine attributes using the same dataset. The detection rate graph for five attributes and nine attributes for anomaly detection is shown in below, and the false-positive rate analysis is shown

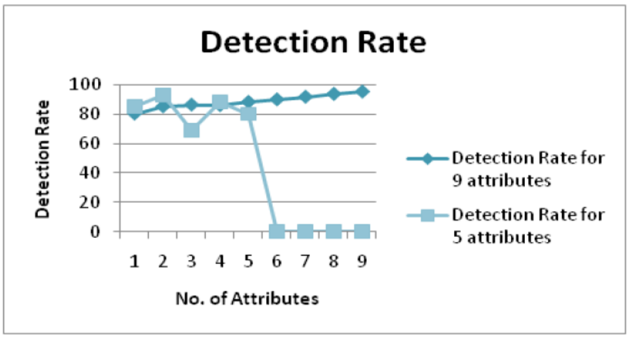


Figure 10 Detection Rate Comparison

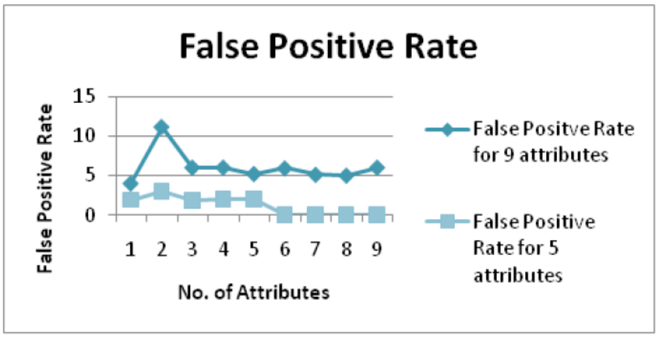


Figure 11. False Positive rate analysis

The Real-time Data performance is compared with KDD CUP 1999 data. The detection rate for normal data using the KDD Cup 1999 Dataset is comparatively lower than the detection rate for Real-time data without considering the attacks. The detection rate is shown in below, and the true positive rate is also high compared to the KDD CUP 1999 dataset. It is shown
Below

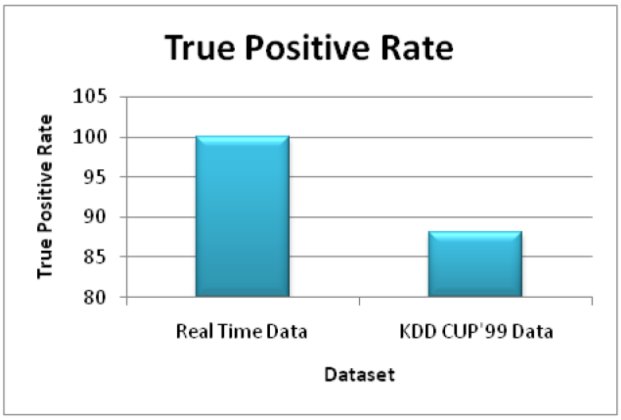


Figure 12. True Positive rate analysis

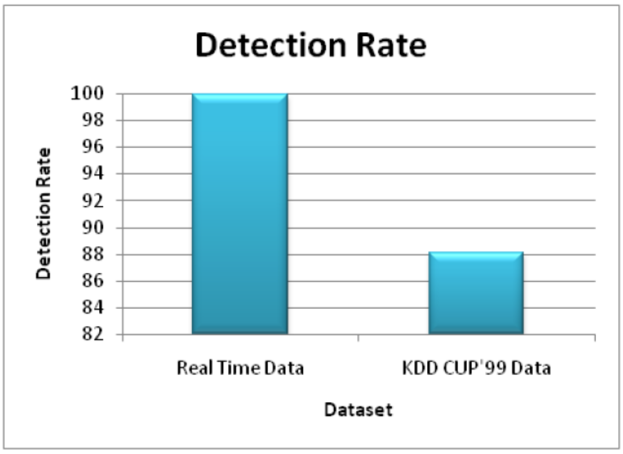


Figure 13. Detection rate analysis

5. Conclusion

This research presents an efficient intrusion detection system using a hybrid approach that combines SVM, FCM, and DBSCAN algorithms. The integration enhances anomaly detection accuracy while reducing false positives. Experimental results confirm the effectiveness of the proposed model in identifying intrusions, making it a viable solution for improving network security.

The KDD CUP dataset is used to classify traffic as normal or DDoS using a Bayesian method. Severe attacks trigger alarms and IP blocking via a block master. Non-DDoS attacks are stored for reference. Performance is evaluated using accuracy, record count, and false positive rate.

The third sturdy is proposed to minimize the resource overhead and results in efficient usage of the available resources. The experimental results shows that Detection rate maximized then the false positives could be reduced. For efficient memory utilization the cascading binary search tree is applied. It helps to economical storage of all data in memory.Thus, the proposed system shows that increased detection rate and minimized false positive rate for testing the data.

For memory efficiency, cascading binary search tree is used where the duplicates or the repeated records are stored only once. Hence, redundancy is avoided. As every record is stored only once, this ensures searching to be more efficient. Memory efficiency could be further enhanced if these data are saved in the form of rules. This will enhance the search methodology.

References

➢ Tartakovsky, A. G. (2006). A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods‟, IEEE Transactions on Signal Processing, vol. 54, no. 9, pp. 3372-3382.

➢ Das, A. (2008). An FPGA-based network intrusion detection architecture‟, IEEE Transactions on Information Forensics and Security, vol. 3, no. 1, pp. 118-132.

➢ Zhao, D. (2010). Research and design for intrusion detection system with hybrid detector and apriori algorith', IEEE 2nd International Conference on E-business and Information System Security, pp. 145-156.

➢ Chen, C.-Y. (2011). Transaction pattern- based anomaly detection algorithm for IP multimedia subsystem', IEEE Transactions on Information Forensics and Security, vol. 6, no. 1, pp. 152-161.

➢ Sperotto, A. (2012). Autonomic parameter tuning ofanomaly-based IDSs: An SSH case study', IEEE Transactions on Network and Service Management, vol. 9, no. 2, pp. 128-141.

➢ Nadeem, A. (20213). A survey of MANET intrusion detection & prevention approaches for network layer attacks‟, IEEE Communications Surveys and Tutorials, vol. 15, no. 4, pp. 2027-2045.

➢ S, A. (2013). Alexander G Tartakovsky, Aleksey S Polunchenko & Grigory Efficient computer network anomaly detection by change point detection methods‟, IEEE Journal of Selected Topics in Signal Processing, vol. 7, no. 1, pp. 4-11.

➢ Sun, B. (2013). Anomaly detection based secure in-network aggregation for wireless sensor networks‟, IEEE Systems Journal, vol. 7, no. 1, pp. 13-25.

➢ Fung, C. J. (2017). Dirichlet-based trust management for effective collaborative intrusion detection networks‟, IEEE Transactions on Network and Service Management, vol. 8, no. 2, pp. 79-91.

➢ Tama, B. A. (2018). TSEIDS: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system‟, IEEE Access, pp. 94497-94507.

➢ Park, E. (2018). Anomaly detection for http using convolutional autoencoders', IEEE Access, vol. 6, pp. 70884-70901.

- Salo, F. (2018). Data mining techniques in intrusion detection systems: A systematic literature review', IEEE Access, pp.56046-56058.

- Benmalek, M. (2024). Advances in Artificial Intelligence and Machine Learning, This paper presents an approach to enhancing the efficiency and effectiveness of NIDS by leveraging Machine Learning techniques, specifically Decision Trees, Naïve Bayes, and Support Vector Machine.

- Zhao, X. (2024). This research proposes a novel approach for enhancing the performance of NIDS through the integration of Generative Adversarial Networks.

- Holdbrook, R. (2024). Electronics, MDPI, This survey explores the specialized requirements, advancements, and challenges unique to deploying NIDS within industrial and robotic systems.

- Kimanzi, R. (2024). The increase in network attacks has necessitated the development of robust and efficient intrusion detection systems (IDS) capable of identifying malicious activities in real-time. In the last five years, deep learning algorithms have emerged as powerful tools in this domain, offering enhanced detection capabilities compared to traditional methods.

- Thiyam, B. (n.d.). *2024*. The evolutions in information and communication technology (ICT) devices have led to the rise in cyber-attacks in network systems. Ensuring cybersecurity in these devices is one of the most critical and complex issues.