e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

NETWORK INTRUSION DETECTION SYSTEM

Aman Balhara[*1], Aman Jaglan[*2], Dr. Arushi Jain[*3], Dr. Sunil Maggu[*4], Dr. Vinay Saini[*5]

[*1,2]Student, Information Technology, Maharaja Agrasen Institute Of Technology, Rohini – Delhi, India.

[*3,4,5]Assistant Professor, Information Technology, Maharaja Agrasen Institute Of Technology, Rohini, Delhi, India.

ABSTRACT

IDS stands for Intrusion Detection System. It helps in monitoring network traffic and detects unwanted intrusions to secure our system from any kind of attacks or malwares. We basically used KNN datasets and NIDS as an algorithm to find the best possible way to get our results. We have used different models like Decision Tree Model, Gaussian Naïve Baye Model,

K-Neighbors Classifier Model, Logistic Regression model. To build an algorithm for processing the data and network and monitor it for the detection of intrusion. In our project we found that the KNN algorithm is fastest and that's why we adopted it.

Keywords: IDS, KNN, KDD Dataset, Tree Model, K-Neighbors Classifier Model.

I. INTRODUCTION

Intrusion Detection System is a software application that detects various network processing while also keeping track of intrusions. The internet has accelerated the development of new application areas. The use of LAN and WAN has risen in a variety of areas, including corporate sectors, industrial regions, security, and healthcare facilities, among others, and this has made the network more appealing and engaging, while also making it vulnerable to attack. KNN algorithm was found to be significant in NIDS[1] among all the data mining techniques tested. This technique, however, necessitates a significant amount of time and storage space. As a result, the fast KNN algorithm was suggested as a solution to this problem. To safeguard the private network, many more firewall mechanisms were devised. The network does not function in isolation; data must be processed throughout the process. Then there's the KDD'99 data collection, which has improved over time by eliminating duplicate entries. Various machine learning methods, such as the SVM approach and decision tree, were used to target this data set with various assaults such as probing and DoS. The goal of intrusion detection is to monitor network assets in order to spot unusual behaviour and network misuse. This notion was proposed in the early 1980s, after the emergence of the internet, with surveillance and threat monitoring[2]. Unexpectedly, notoriety and inclusion into security infrastructure grew. Since then, several events in IDS technology have led to the current level of high-end intrusion detection. Cisco saw network intrusion detection as a priority and acquired the Wheel group to provide securitysolutions.

In the 1990s, a significant quantity of income was earned, and the intrusion detection business grew. Real secure is an intrusion detection system network built by ISS. In a report for a government entity, James Anderson proposed that audit trails hold highly essential information that may be useful in detecting misuse and gaining a better knowledge of user behaviour. Host-based IDS, Network-based IDS, and Application-based IDS arethe three forms of IDS.

The audit trail is very important to a host-based system. The data enables the intrusion detection system to discover minor patterns of misuse that would otherwise go undetected at a higher level of abstraction. Sensor refers to a host-based handler. The fundamental premise of IDS, especially Network Based Intrusion Detection System (NIDS), stems from Denning's pioneering work on anomalous HIDS. HIDS are effective for assessing network assaults because they can identify exactly what the attacker did, which commands he used, and which files he opened, rather than merely making a broad charge and attempting to perform a harmful operation. Configuration is the safest option. Advantages of Intrusion Detection Systems Based on the Host Reduced entry costs, reduced entry costs, and real-time detection and reaction.

Instead of collecting data from each individual host, network-based IDS systems collect data from the network itself. The sensor attack is based on signatures that are similar to earlier assaults, and the functioning of the

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

monitors will be transparent to the users, which is also important. NNIDS (Network Node IDS) agents are installed on each host in the network being protected. Attack signatures are rules that describe what constitutes an attack in network sensors, and most network-based systems allow sophisticated users to specify their own signatures.

Network-based Intrusion Detection Systems have the following advantages: they are easier to implement, they retain evidence, they detect unsuccessful assaults, they have a lower cost of ownership, and they detect network-based attacks.

The protocol's effective behaviour and events will be checked by application-based IDS (APIDS). Unintentional assaults inflict financial damage to the company by destroying key data files. Intentional attacks are malicious attacks carried out by dissatisfied workers to bring harm to the firm. The agent or system is positioned between a process and a collection of servers in this case, and it watches and analyses the application protocols between devices. Application-based Intrusion Detection Systems provide the following advantages: Application-Based IDS tracks activities to individual users by monitoring the interaction between the user and the application. Because it communicates with the application at transaction endpoints where information is exposed to users in an unencrypted form, application-based IDS works with apps that access encrypted data.

Attacks :

- **CYBER ATTACKS**

A cyber attack is a cybercriminal attack that uses one or more computers to attack a single or numerous computers or networks[3].

- **Denial of service attacks (DOS)**

It's the most damaging type of cyber-attack since it works by generating a lot of traffic in the memory or computational resource. Back, Neptune, Land, Teardrop, Smurf, and TCP SYN flooding are all examples of this assault.

- **Remote to local (User) Attack (R2L)**

Ftp-Write, Xsnoop, Guest, and the Dictionary are examples of attacks that target misconfigured or poor system security. Another assault that employs social engineering to acquire access is the Xlock attack.

This type of attack sends packets to the network with the goal of exploiting vulnerabilities in order to get unauthorised local access to network resources.

- **User to Root Attacks (U2R)**

This lesson begins with acquiring access to a normal user and sniffing for passwords in order to acquire access to computer resources as a root user. The most prevalent U2R attack is buffer overrun.

- **Probing**

They frequently get privileged access to a non-expecting host by exploiting a known vulnerability. Probing is a type of attack in which an attacker scans a network for flaws such as open ports that may be exploited to identify services running on the resource.

TOOLS :

A. KDD'99 CUP DATASET

KDD CUP 99 in-depth analysis

The KDD'99 data set has been the most widely used data set for testing anomaly detection systems since 1999. This data collection was created by Stolfo et al. It is based on the data collected during DARPA's IDS evaluation programme in 1998, and it serves as a standard against which to compare alternative approaches.

DARPA 98 consists of around 4 gigabytes of compressed raw tcp dump data from 7 weeks of network activity, which may be processed into approximately 5 million connection records, each with approximately 100 bytes. It provides a training dataset of roughly 4,900,000 single connection vectors, each of which has 41 characteristics and is labelled as either normal or attack, with only one attack type. Denial of service (dos), user to root (u2r), remote to local (r2l), and probing assaults are the four types of attacks that may be found in a military network.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

Denial of Service Attack (DoS): When an attacker tries to prohibit legitimate users from utilising a service, this is known as a DoS attack. This is the most serious type of cyber-attack, in which a large amount of traffic is generated within a computational or memory resource, causing it to become overburdened and unable to process requests from authorised users of the system. User to Root Attack (U2R): This is a type of exploit in which the attacker gets local access to the victim's computer and attempts to gain superuser capabilities. An attacker who has the ability to transmit packets to a system across a network commits a remote to local attack (R2L). Probing Attack: In this attack, the attacker attempts to obtain vital information about the computer network (target host).

Using the Bro IDS, the KDD dataset was completed,

yielding 41 features for each and every connection.

There are four different types of features:

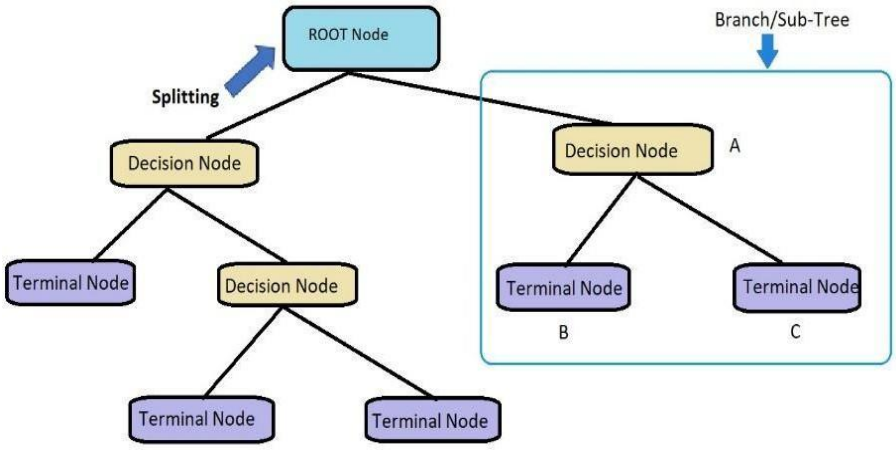
Basic characteristics are those that are extracted from packet headers without examining the payload.

Time-based Traffic Features: It is intended to capture qualities that develop over a two-second period of time. Features such as the number of failed login attempts are included in the content. The payload of the original TCP packets is accessed using domain knowledge.

Host-based Traffic Features: Instead of using time, use a historical window calculated over the number of connections, in this example 100 instead of time.

B. DECISION TREE

It primarily consists of a two-stage process, with the first phase being learning and the second being prediction. The most frequent algorithm for interpreting in the simplest method is the decision tree. It's used to hone the model's ability to forecast class/value. The categorical variable decision tree is one of two types of decision trees. One is a categorical variable decision tree, while the other is a continuous variable decision tree, which has a continuous target variable. Figure 1 depicts the decision tree method in its most basic form.



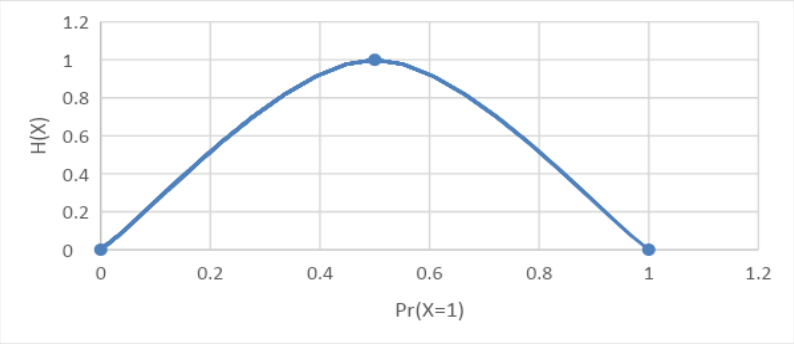
Decision trees utilise a variety of strategies to divide nodes into more. There is a direct link between the target variable and node purity. The nodes in a decision tree are divided based on the variables that are accessible. The following are some of the algorithms that are utilised in decision trees: CHAID, ID3, C4.5, CART, MARS.

C. ATTRIBUTE SELECTION MEASURES

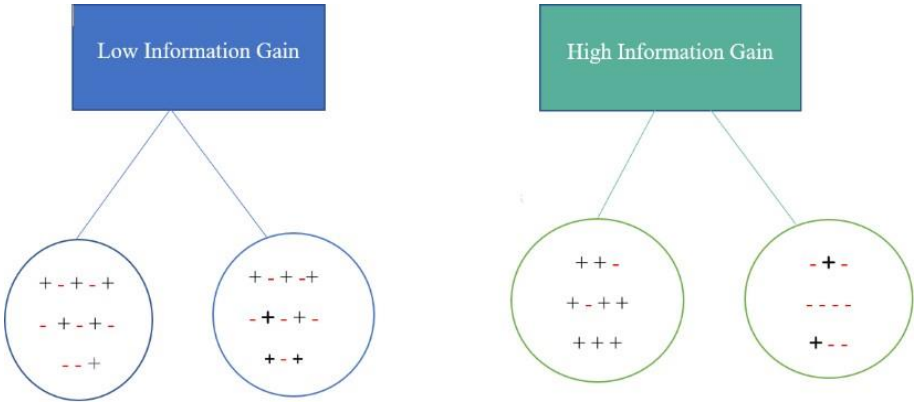
Because randomly picking any node for different levels compounds the challenge, researchers attempted to find methods that were both simple and effective.

ENTROPY- As seen in Figure 2, entropy is a measurement of the unpredictability of occurrences. H(X) represents entropy, whereas Pr(X) represents probability in the network. It also shows that entropy is greatest when probability is precisely half, which occurs since there is no ideal method for calculating data randomness.

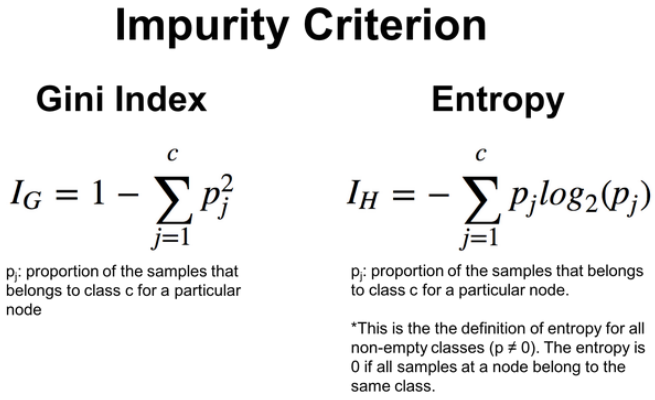
e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com



INFORMATION GAIN - A statistical feature is required to separate training samples by provided attributes according to the corresponding goal categories, which is defined as Information gain. Figure 3 elucidates the above-mentioned terms. It is a decrease in entropy that the ID3 decision tree method employs. IG may be expressed mathematically as:



GINI INDEX - This is a cost function that is used to evaluate the splits in a dataset. The computation is done by removing one from the total of each class's squared probability. It is thought to be superior to the computation of knowledge gain. The fundamental comparison between the two is shown in Figure 4.



GAIN RATIO - Information gain is a gain ratio change that lowers bias and is the best alternative. In reality, the gain ratio is used to adjust IG by using intrinsic information. The formula for calculating the gain ratio is given below.

K= number of subsets generated by the split

$$Gain\ Ratio = \frac{Information\ Gain}{SplitInfo} = \frac{Entropy\ (before) - \sum_{j=1}^{K} Entropy(j, after)}{\sum_{j=1}^{K} w_j\ log_2\ w_j}$$

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

CHI-SQUARE - The chi-squared Automatic Interaction Detector, or CHAID, is one of the earliest tree categorization algorithms. The category goal variable "SUCCESS" or "FAILURE" is used. The statistical significance is exactly proportional to the chi-squared value. The chi-square split formula is shown in Figure 5.

$$\chi^2 = \sum_{j=1}^{k} \frac{(f_{b_j} - f_{e_j})^2}{f_{e_j}}$$

D. MODELS

There are a few models that are utilised, and they are as follows:[8]-

Model of the K-Neighbor Classifier. Model of Logistic Regression

Model of Gaussian Naive Bayes

II. COMPARISON BETWEEN ACCURACY

Accuracy is a metric used in machine learning to determine how often a model correctly predicts data. The K-Nearest Neighbor classifier in figure 9 has the best accuracy of all the models, while the decision tree model in figure 2 has the lowest. While the accuracy of the naive bayes classifier in figure 1 and logistic regression in figure 3 are comparable to that of the K-Neighbor and decision tree, respectively.

```
Model Accuracy:
 0.8165880365775525

Confusion matrix:
 [[5591 1867]
 [1282 8429]]

Classification report:
              precision    recall  f1-score   support

         0.0       0.81      0.75      0.78      7458
         1.0       0.82      0.87      0.84      9711

    accuracy                           0.82     17169
   macro avg       0.82      0.81      0.81     17169
weighted avg       0.82      0.82      0.82     17169
```

Figure 1- Naïve bayes model

```
Model Accuracy:
 0.8336536781408352

Confusion matrix:
 [[5487 1971]
 [ 885 8826]]

Classification report:
              precision    recall  f1-score   support

         0.0       0.86      0.74      0.79      7458
         1.0       0.82      0.91      0.86      9711

    accuracy                           0.83     17169
   macro avg       0.84      0.82      0.83     17169
weighted avg       0.84      0.83      0.83     17169
```

Figure 2- Decision tree model

```
Model Accuracy:
 0.8418661541149747

Confusion matrix:
 [[5963 1495]
 [1220 8491]]

Classification report:
              precision    recall  f1-score   support

         0.0       0.83      0.80      0.81      7458
         1.0       0.85      0.87      0.86      9711

    accuracy                           0.84     17169
   macro avg       0.84      0.84      0.84     17169
weighted avg       0.84      0.84      0.84     17169
```

Figure 3- Logistic regression model

```
Model Accuracy:
 0.8666200710583027

Confusion matrix:
 [[5787 1671]
 [ 619 9092]]

Classification report:
              precision    recall  f1-score   support

         0.0       0.90      0.78      0.83      7458
         1.0       0.84      0.94      0.89      9711

    accuracy                           0.87     17169
   macro avg       0.87      0.86      0.86     17169
weighted avg       0.87      0.87      0.86     17169
```

Figure 4– K-Neighbor model

III. EVALUATION AND RESULTS

A confusion matrix is used to assess a classification model's performance. The matrix compares the actual value to the categorization model's anticipated value. A heatmap is a two-dimensional depiction of basic and complicated data using colours. With the aid of a heatmap, the user may readily see both basic and complicated information. Figure 5 depicts a heatmap of a decision tree classifier's confusion matrix, while Figure 6 depicts the confusion matrix of a logistic regression using heatmap. Figures 7 and 8 show the confusion matrix of naive bayes and k-neighbor classifiers as a heatmap.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

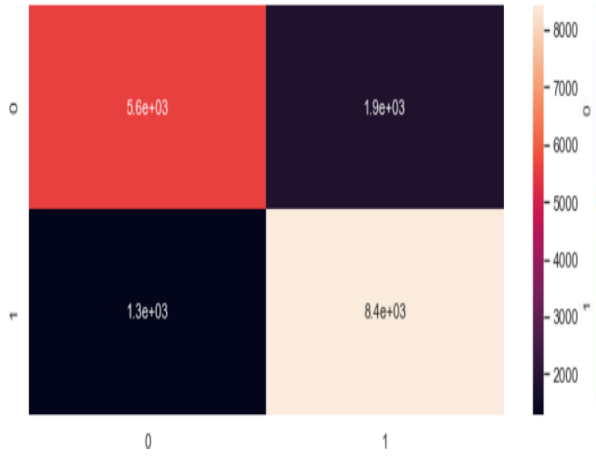


Figure 5–Decision tree heat map

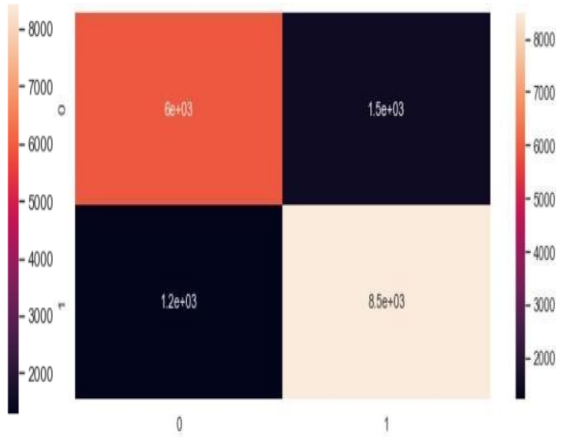


Figure 6- Logistic regression heatmap

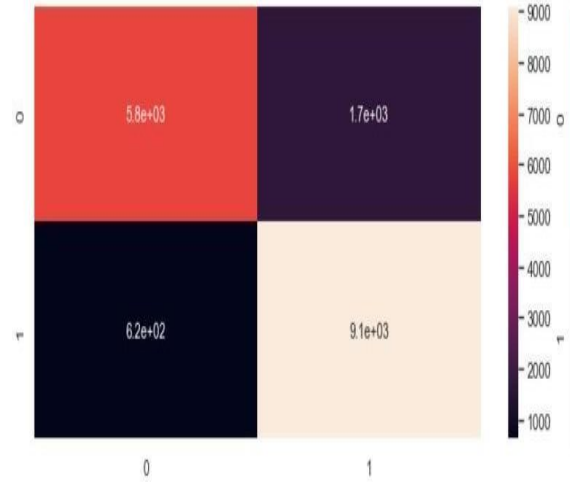


Figure 7– naïve bayes model heat map

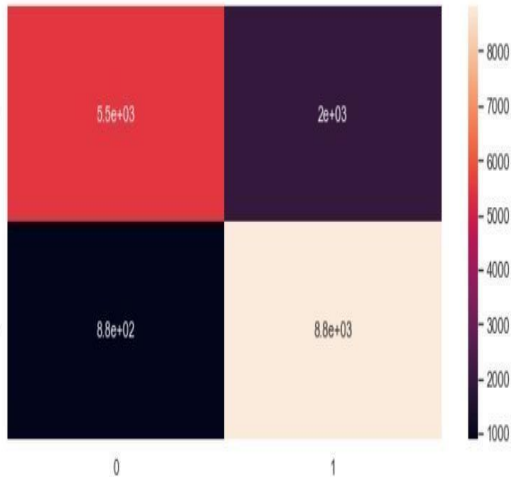


Figure 8- K-Neighbor model heatmap

MODELS	ACCURACY	PRECISION	RECALL	F1-SCORE
Decision Tree	0.8165	0.82	0.82	0.82
K-Neighbor	0.8666	0.87	0.87	0.86
Logistic Regression	0.8418	0.84	0.84	0.84
Naïve Bayes	0.8336	0.84	0.83	0.83

IV. CONCLUSION

There is yet to be developed in the field of Network Intrusion Detection System. With the use of the KDD dataset, NIDS is implemented utilising several models such as the K-Neighbor model, decision tree model, logistic regression model, and naive bayes model. It is critical to use NIDS to detect network threats and to prevent network infiltration. This necessitates the implementation of a network intrusion detection system (NIDS) that reliably alerts the user of any network threats. As a result, while developing a model for a dataset including several network assaults, we discovered that including more characteristics in the dataset improved the model's overall accuracy. Furthermore, the use of a Neural Network model to a dataset containing values has been shown to be extremely successful in classifying all types of assaults. As a result, K-Neighbor is the best algorithm for developing a Network Intrusion Detection system, whereas Decision Tree has the lowest accuracy of all.

V. FUTURE WORK

Currently, the suggested model only recognises four primary types of assaults. There will be many more assaults in the future, and the system will always be susceptible. As a result, engineers must search for any

conceivable gaps that might compromise data security. Developing an NIDS that properly detects all network assaults is a major problem. As a result, we've used a deep learning method to optimise the system and make it safer.

VI. REFERENCES

[1] Selecting features for intrusion detection: A feature relevant analysis on KDD 99 Intrusion Detection Datasets

[2] https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained

[3] An efficient classifier for U2R,R2L,DoS ATTACK, IJRTE, Piyush Gupta

[4] Two feature selection algorithms based on ensemble of SVM Classifier for Intrusion Detection.

[5] Peyman Kabiri and Ali A.Ghorbani-"Research on Intrusion and Response Survey"-International Journal of Network Security.

[6] Christopher Low – "understanding wireless attacks & Detection – GIAC Security essentials certification practical assignments.

[7] Fast kNN Classifier for network intrusion detection system, B.baseswara Rao and K.Swathi.

[8] https://www.ripublication.com/ijaer19/ijaerv14n5_10.pdf

[9] Intrusion detection system- A study, Dr. S.Vijayarani and Ms. Maria.