IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, May 2022
Impact Factor: 6.252

# Network Intrusion Detection by Machine Learning Techniques

**Keshav A Ruia[1], Milap M Alondra[2], Sahil M Shrivatsava[3], Mitesh V Salunke[4], Dr. Geeta S Navale[5], Mrs. Supriya H Lokhande[6]**
Students, Department of Computer Science Engineering[1,2,3,4]
Faculty, Department of Computer Science Engineering[5-6]
Sinhgad Institute of Technology and Science, Pune, Maharashtra, India

**Abstract:** *The web has been utilized broadly in all parts of life. The Interference of web associations can create a huge effect. Hence, the job of the Network Intrusion Detection System (IDS) to distinguish digital attacks is vital. A suspicious connection needs to be blocked immediately before performing anything further. The Higher the data transmissions occuring daily its being important to protect the data and its been main factor to prevent intrusions. A good Intrusion System is to be developed to prevent Attacks. This paper presents a novel approach to classify intrusion attacks. The focal thought is to apply different machine learning algorithms like SVM, Naive Bayes, Neural Networks, Random Forest, Logistic Regression. We apply these kinds of supervised and unsupervised learning Techniques and classify the attack classes. The presentation of the various models was analyzed utilizing every one of the highlights and the best-chosen highlights were executed utilizing the disarray grids.*

**Keywords:** Feature Extraction, Deep learning, ML, Transfer Learning, KDD, Train Set, Test Set

## I. INTRODUCTION

Nowadays, frameworks that are connected to networks through the web, which are continually powerless against networking attacks. It is critical to diagnose intruders as soon as possible in order to carry out the necessary preventative measures. Because of the variety of attack forms.

Attackers take advantage of the available flaws such as insufficient security procedures and practices, as well as computer defects such as system vulnerabilities, resulting in network breaches. The attackers might be less privileged device operators looking to get more access control, or black hat-hackers looking to steal sensitive information from everyday internet users [1]. Methods for detection can be founded on distinguishing usage or in view of exception identification.

An attack can occur as a result of user error, system configuration errors, or campaign flaws. In millions of computers linked to the internet network, there is always the possibility of an intruder. Because it is difficult to detect an attacker based on aspects of the hypervisor, IP address, application, or hardware used, an automated infrastructure is likely to overcome this situation.

IDS can be partitioned into two sorts as indicated by the strategies for their development, which are host-based intrusion detection systems and network-based intrusion detection systems. Network-Based NIDS are distributed at strategic points to monitor packet transmission in the network as it comes and goes to all network devices. This data monitoring is then assessed and compared to known attack signatures. It is simple and inexpensive to utilise on the network's edge, that it can supervise all traffic.

- Host-Based HIDS is accomplished on individual network host systems and monitors all data flow rather than the entire network. If somehow the compromised node is still in the network, HIDS is more useful.

| Attack Category | Examples |
|---|---|
| Dos | teardrop ,Back, neptune, smurf |
| Probe | Satan, portsweep, ipsweep, nmap |
| R2L | spy Guess password, imap, mutihop |
| U2R | perl.rootkit, buffer overflow, load module, |

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, May 2022
Impact Factor: 6.252

The recreated assaults can fall in any of the accompanying four classes.

- **Probing Attack**: This is a sort of attack which gather data of target framework before starting an assault. A piece of the models are Satan, ipsweep, nmap attacks.
- **DoS Attack:** Denial of Service (DoS) attack results by forestalling authentic solicitations to an organization asset by consuming the transmission capacity or by over-burdening computational assets. Instances of this are Smurf, Neptune,
- **Client to Root (U2R) Attack:** For this situation, an attacker begins with admittance to an ordinary client account on the framework and can take advantage of the framework weaknesses to acquire root admittance to the framework. Models are discharge, load module and Perl assaults.
- **Root to Local (R2L) Attack**: In this, an aggressor who doesn't have a record on a remote machine sends parcel to that machine over an organization and takes advantage of certain weaknesses to acquire neighborhood access as a client of that machine. A few models are ftp_write, surmise secret key and imap assaults

## II. EXISTING SYSTEM

In [3] the discovery of interlopers in light of irregular woodlands and k-implies was proposed. The arbitrary backwoods is utilized as a classifier to make assault designs. Your strategy utilizes the KDD99 record. Hesham alwaijry created IDS involving Bayesian likelihood in his article [4]. Their technique further fostered the area rate to 85.35 for R2L attacks. The makers used the KDD dataset for their preliminary examination. In [5] the creators proposed IDS in light of SVM. Their methodology comprises of three stages: 1) preprocessing: used to preprocess TCP/IP information, 2) preparing: preparing the information to identify assaults, and 3) testing: estimating the exhibition. . Hansung Lee et.on the proposed IDS, which maintains the benefits of both abnormality based and signature-based IDSs. Its IDS is a multi-pass, multi-class IDS that speeds up [6]. In Ref. [7] the creators propose an effective information fitted choice tree for IDS. His proposed approach checked out at the various sorts of assaults. A GA-based interference acknowledgment structure is proposed by Dheeraj et.to [8]. The creators proposed a fluffy quality IDS coordinated with characteristic choice. Their methodology accomplished high location rates and low phony problem rates.

## III. METHODOLOGY

It Involves several processes as we describe with the use of Datasets then we use feature extraction, which involves transfer learning and then we apply classification techniques. Lastly we use several metrics to compare its approaches

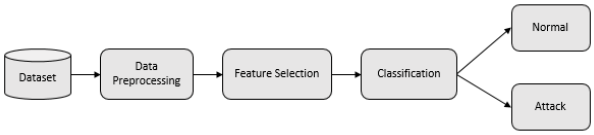



**Figure 1:** System architecture

### 3.1 Dataset

KDDCup 1999 is a standard dataset for interruption identification. KDDCup 1999 dataset The principal huge impediment inside the KDDCup 99 dataset is the enormous number of repetitive records, with almost 78% of preparing and 75 percent of testing records being copied. This causes the order model to be one-sided on the most incessant records, keeping it from perceiving uncommon assault documents that fall into the U2R and R2L classes.

Simultaneously, it initiates the summarizing to be slanted by strategies with higher discovery rates on continuous events. We made two datasets to all the more likely assess the proposed strategy: Dataset An and Set of information B. The kdd boundaries are something very similar in both datasets. The classes are adjusted in both datasets. During the readiness stage, the class is included a wide range of sources. Both of these channels were gotten to on March 31, 2020. Coming up next are the advantages of NSL KDD over KDD 99: It has a lower skewed esteem in light of the fact that no excess information is incorporated, and the quantity of records picked is relative. The informational index contains 38 assault types partitioned into four classes: DoS, Probe, U2R and R2L. The appropriation of the preparation informational index comprises of 53.4% typical associations, 36.4 DoS assaults, 9.25% test assaults, 0.79% of R2L assaults and 0.04% of U2R assaults; Test dataset

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, May 2022
Impact Factor: 6.252

incorporates 43% ordinary associations, 33% DoS assaults, 10.7% examining assaults, 12.2% R2L assaults, and 0.89% U2R assaults

### 3.2 Feature Extraction Steps

Its is one of the normal terms utilized in information mining to reduce the input to which the data can be analyzed and processed. A few strategies are accessible that are utilized to characterize the elements. Filter Method This method is based on the selection of features using statistical measurements Each feature is scored based on statistical calculations to determine the threshold value depending on the intrinsic properties of the data. Bagging Algorithms methods include analysis hypothesis within the subset of features wherein a search is determined in the subset space of potential features, also, different subsets of elements are made and assessed.

### 3.3 Architectures Used

#### A. K-Nearest Neighbor Algorithm (KNN)

The K-Nearest Neighbor Algorithm is a methodology that applies to Regression and Classification. In both regression and classification, the k-entry includes the most private training events throughout the domain. Subsequently, the result depends on the utilization of KNN. The KNN algorithm is used to find the k-samples in the training set (k is a positive integer, usually small). The training samples are vectors in different dimensions; moreover, it helps to select the predominant class of training K-samples. This class is allocated as a prevalent for the objective example, where k is the quantity of preparing tests.

#### B. Decision Tree Algorithm (DF)

A Decision tree is drawn with its root at the top.Decision Tree has a place with the class of the trees and it comprises of the root hub then a few branches goes through the root hub.Several children nodes are formed as final node to the result.

#### C. Support Vector Machine Algorithm (SVM)

SVM makes an optimization level by isolating data from different categories. The data are modified for several classes. Based on this separation, the complexity Increases

#### D. Naive Bayes

Bayes classifiers are a group of straightforward "probabilistic classifiers" in view of applying Bayes'-hypothesis for multi-class characterization problems. Naive Bayes can be viewed as a regularized type of the Bayesian order system by limiting the covariance framework to be slanting.

$$P(c|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|c)P(c)}{P(\boldsymbol{x})},$$

#### E. Logistic Regression

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability .Strategic Regression a Linear Regression model however the Logistic Regression utilizes a more intricate expense work, this cost capacity can be characterized as the 'Sigmoid capacity' or otherwise called the 'calculated work'.

#### F. Neural Network

A Neural network classifier comprises of number of neurons units, organized in layers. Each layer takes some information vector and gives yield by applying a non-direct capacity. This result functions as contribution to next layer in feed-forward way. As a general rule, there is no input to past layer. Last result layer has execution of grouping of assault.

### 3.4 Performance Metrics

The Evaluation of this model has been carried out with help of various metrics. The performance metrics derived from the confusion matrix has been utilized as a measurements for the assessment of this models. The disarray lattice gives a four

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, May 2022
Impact Factor: 6.252

results of misleading negative (FN), bogus positive (FP), genuine negative (TN), and genuine positive (TP). The presence of both FN's and FP's could affect decisions negatively. An FP result is occurred when the wrong attack is predicted in case of proper detection. An FN occurs when an individual who is supposed to fall into a given class is instead excluded from this group. The performance of the different networks was evaluated on the test set by computing the macro average of accuracy (Acc), score, precision (PPV), specificity (Spc).

$$\text{Accuracy (Acc)}_i = \frac{TP_i + TN_i}{TP_i + FP_i + TN_i + FN_i}, \quad (1)$$

$$F1\ \text{score}_i = 2 \times \frac{PPV_i \times Sen_i}{PPV_i + Sen_i}, \quad (2)$$

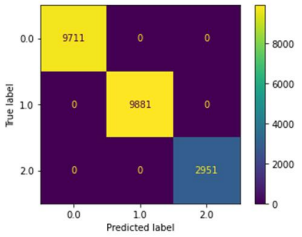
$$\text{Precision (PPV)}_i = \frac{TP_i}{TP_i + FP_i}, \quad (3)$$

$$\text{Specificity (Spc)}_i = \frac{TN_i}{FP_i + TN_i}, \quad (4)$$

## IV. Evaluation and Results

Following are a few important results and plots that help estimate the accuracy of the models and get insights their performance.

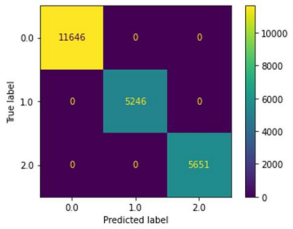
### 4.1 Logistic Regression




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 1.00 | 1.00 | 1.00 | 9711 |
| 1.0 | 0.73 | 0.96 | 0.83 | 7459 |
| 2.0 | 0.68 | 0.83 | 0.74 | 2421 |
| 3.0 | 0.00 | 0.00 | 0.00 | 2885 |
| 4.0 | 0.00 | 0.00 | 0.00 | 67 |
| accuracy | | | 0.84 | 22543 |
| macro avg | 0.48 | 0.56 | 0.51 | 22543 |
| weighted avg | 0.74 | 0.84 | 0.78 | 22543 |

the accuracy= 0.8376879740939538

### 4.2 K-Nearest Neighbor

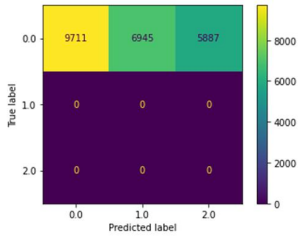


IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, May 2022
Impact Factor: 6.252

| | | | | |
|---|---|---|---|---|
| 0.0 | 0.69 | 0.83 | 0.75 | 9711 |
| 1.0 | 0.77 | 0.54 | 0.64 | 7459 |
| 2.0 | 0.28 | 0.65 | 0.39 | 2421 |
| 3.0 | 0.00 | 0.00 | 0.00 | 2885 |
| 4.0 | 0.00 | 0.00 | 0.00 | 67 |
| accuracy | | | 0.61 | 22543 |
| macro avg | 0.35 | 0.40 | 0.36 | 22543 |
| weighted avg | 0.58 | 0.61 | 0.58 | 22543 |

the accuracy= 0.6059087078028657

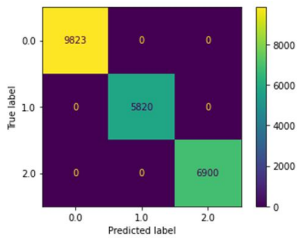
### 4.3 Decision Tree




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.43 | 1.00 | 0.60 | 9711 |
| 1.0 | 0.00 | 0.00 | 0.00 | 7459 |
| 2.0 | 0.00 | 0.00 | 0.00 | 2421 |
| 3.0 | 0.00 | 0.00 | 0.00 | 2885 |
| 4.0 | 0.00 | 0.00 | 0.00 | 67 |
| accuracy | | | 0.43 | 22543 |
| macro avg | 0.09 | 0.20 | 0.12 | 22543 |
| weighted avg | 0.19 | 0.43 | 0.26 | 22543 |

the accuracy= 0.4307767377900013

### 4.4 Naive Bayes

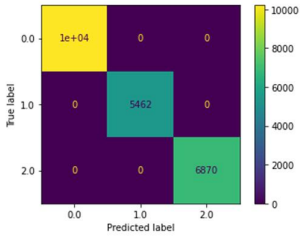



| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.99 | 1.00 | 0.99 | 9711 |
| 1.0 | 0.96 | 0.75 | 0.85 | 7459 |
| 2.0 | 0.32 | 0.92 | 0.48 | 2421 |
| 3.0 | 0.00 | 0.00 | 0.00 | 2885 |
| 4.0 | 0.00 | 0.00 | 0.00 | 67 |
| accuracy | | | 0.78 | 22543 |
| macro avg | 0.46 | 0.53 | 0.46 | 22543 |
| weighted avg | 0.78 | 0.78 | 0.76 | 22543 |

the accuracy= 0.778379097724349

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, May 2022
Impact Factor: 6.252

### 4.5 SVM




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.87 | 0.91 | 0.89 | 9711 |
| 1.0 | 0.89 | 0.65 | 0.76 | 7459 |
| 2.0 | 0.33 | 0.95 | 0.49 | 2421 |
| 3.0 | 0.00 | 0.00 | 0.00 | 2885 |
| 4.0 | 0.00 | 0.00 | 0.00 | 67 |
| accuracy | | | 0.71 | 22543 |
| macro avg | 0.42 | 0.50 | 0.43 | 22543 |
| weighted avg | 0.70 | 0.71 | 0.69 | 22543 |

the accuracy= 0.7109524020760325

### 4.6 CNN

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 1.00 | 1.00 | 1.00 | 9711 |
| 1.0 | 0.92 | 0.90 | 0.91 | 7459 |
| 2.0 | 0.41 | 0.93 | 0.56 | 2421 |
| 3.0 | 0.00 | 0.00 | 0.00 | 2885 |
| 4.0 | 0.00 | 0.00 | 0.00 | 67 |
| accuracy | | | 0.83 | 22543 |
| macro avg | 0.47 | 0.57 | 0.50 | 22543 |
| weighted avg | 0.78 | 0.83 | 0.79 | 22543 |

## V. CONCLUSION

In this work, we explored different avenues regarding numerous AI models trying to characterize the attacks .The KDD dataset was utilized as benchmark to recognize ordinary and strange organization traffic designs. The most associated highlights were separated utilizing factual techniques and were the contribution of a The attainability and viability of the proposed model were assessed utilizing accuracy, review, F measure and exactness measurements. The similar assessment of proposed with a few classifier and cutting edge models showed that the 83% precision for the specific Logitic Regression ML model with most elevated exactness.

## REFERENCES

**[1].** H. Wang, J.Gu, and S.Wang, ‘‘An effective intrusion detection framework based on SVM with feature augmentation,’’ Knowl.-Based Syst., vol. 136, pp. 130–139, Nov. 2021.

**[2].** Setareh Roshan, Yoan Miche, Anton Akusok, Amaury Lendasse; “Adaptive and Online Network Intrusion Detection System using Clustering and Extreme Learning Machines”, ELSEVIER, Journal of the Franklin Institute, Volume.355, Issue 4,March 2018,pp.1752-1779.

**[3].** Reda M.Elbasiony,”A hybrid NIDS framework based on RF and weighted k means”,ali shams engineering journal,pp 753- 762(2020)

**[4].** Hesham Alwaijry,”Bayesian based IDS”,CIS, Journal of king saud university,pp1-6,(2020)

**[5].** Mukkamala,”Intrusion detection using NN and SVM”,IJCNN2002,Vol 2,IEEE(2021)

**[6].** Lee Hansung,”IDS based on multi class SVM” Rough sets, data mining, and granular computing, Springer,pp 511-519

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, May 2022
Impact Factor: 6.252

**[7].** G V Nadiammai,"Effective approach toward IDS using data mining techniques", Egyptian informatics journal,pp37-50

**[8].** Dheeraj pal,"Improved genetic algorithm for IDS",In. sixth international conference on CCICN2014,PP835-839

**[9].** Wathiq Laftah Al-Yaseen , Zulaiha Ali Othman , Mohd Zakree Ahmad Nazri; "Multi-Level Hybrid Support Vector Machine and Extreme Learning Machine Based on Modified K-means for Intrusion Detection System", ELSEVIER, Expert System with Applications,Volume.66,Jan 2020,pp.296-303.

**[10].** Iftikhar Ahmad, Mohammad Basheri, Muhammad Javed Iqbal, Aneel Raheem; "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection", IEEE ACCESS, Survivability Strategies for Emerging Wireless Networks, Volume.6,May 2020,pp.33789-33795.

**[11].** BuseGulAtli1, Yoan Miche, Aapo Kalliola, Ian Oliver, Silke Holtmanns, Amaury Lendasse; "Anomaly-Based Intrusion Detection Using Extreme Learning Machine and Aggregation of Network Traffic Statistics in Probability Space" SPRINGER, Cognitive Computation, June 2020,pp.1-16

**[12].** Pinjia He, Jieming Zhu, Shilin He, Jian Li, and Michael R. Lyu; "A Feature Reduced Intrusion Detection System Using ANN Classifier", ELSEVIER, Expert Systems with Applications,Vol.88,December 2021pp.249-247

**[13].** Vajiheh Hajisalem, Shahram Babaie; "A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection", ELSEVIER, Department of Computer Engineering, Vol. 136, pp. 37-50, May 2021.

**[14].** R. G. Bace, "Intrusion Detection", Macmillan Technical Publishing. 2021

**[15].** Freund,Y mason L, "The alternating Decision tree learning algorithm", In proc of 6th ICM Bled, Slovenia,pp 124-133

**[16].** M. A. Jabbar," Alternating decision trees for early diagnosis of heart disease", Proceedings of International Conference on Circuits, Communication, Control and Computing ,pp 322-328

**[17].** L. Dhanabal and D. S. P. Shantharajah, "A Study On NSL-KDD Dataset For Intrusion Detection System Based On Classification Algorithms," Int. J. Adv. Res. Comput. Commun. Eng., vol. 4, no. 6, pp. 446–452.