ASSESSING MACHINE LEARNING TECHNIQUES FOR INTRUSION DETECTION: A COMPARATIVE STUDY

#1KANDUKURI CHANDRASENA CHARY, *Research Scholar,*
#2Dr. SATISH NARAYAN GUJAR, *Supervisor*
Department of Computer Science and Engineering,
School of Engineering and Technology,
UNIVERSITY OF TECHNOLOGY, JAIPUR, RAJASTHAN.

ABSTRACT: It is necessary to investigate network traffic classification using machine learning since the internet is becoming into a more global medium for information sharing. Security flaws impact not just individuals but also entire companies. This makes it crucial to distinguish between reliable and unreliable information on the network. This research will analyze and compare seven distinct machine learning techniques. These include C4.5, XGBoost, Random Forest, Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), and k-Nearest Neighbors (KNN). These investigations operate seamlessly and programmatically thanks to the use of Python's package module. Recall, accuracy, and precision are some of the metrics that are examined in the evaluation since they provide important insights into the effectiveness of each technique.

Keywords: Network Traffic Classification, Machine Learning, KNN, SVM.

1. INTRODUCTION

Standard intrusion detection methods are currently incapable of comprehending the most recent cyber events and threats. The conventional approach, which entails manually analyzing networks or establishing patterns that are illogical, may not be capable of identifying extensive attacks. The internet facilitates the flow of a greater volume of network data, which complicates the task of network analyzers in identifying intrusions due to the ease with which policy information can be accessed.

In order to automate the process of identifying intrusions, it is necessary to employ dynamic and effective methods that can identify and detect novel types of intrusions. This investigation introduces novel intrusion detection methodologies that are highly adaptable and dynamic, and are designed to manage a substantial volume of network data.

The three most critical stages in the process of identifying intrusions are the following: the definition and extraction of features, the definition and extraction of rules, and the application of these rules to identify intrusions in the dataset. These methods are designed to accommodate the distinctive requirements of a variety of networks and systems. Numerous specialists have devised various methods to efficiently organize network data into distinct categories over the past three decades. We discuss a few of the methods that specialists have employed in the past to classify networks. In the past, various methods, such as Classification Based on Port Number and Classification Based on Payload, were employed to organize network data into distinct categories.

2. RELATED WORK

The authors employed the KDD Cup 99 dataset to differentiate between normal and aberrant data. In 2017, Jamal H. Assi and Ahmed T. Sadiq employed the NSL-KDD dataset in a separate study to categorize network risks into distinct categories. C4.5, Bayesian Network, Back Propagation Neural Network, Support Vector Machine (SVM), and Decision Table (DT) were among the classification methods employed in this investigation.

Additionally, numerous feature selection methodologies were implemented, including Decision Tables, Information Gain (IG), and Correlation-based feature selection (CFS). It is crucial to bear in mind that the C4.5 classification method, which incorporates information gain feature selection, outperformed the other methods. Nabin Kumar Karn, Asif Ali Laghari, Lu Yao, Muhammad Shafiq, Xiangzhan Yu, and others devised a method for organizing network data into distinct categories.

This method employed four machine learning algorithms: C4.5, Support Vector Machine, BayesNet, and NaïveBayes, in addition to supervised learning methods. Dhanabal and Shantharajah (2015) employed the NSL-KDD dataset to evaluate various classification methods, including SVM, Naïve Bayes, and J48, with an emphasis on network packet defects.

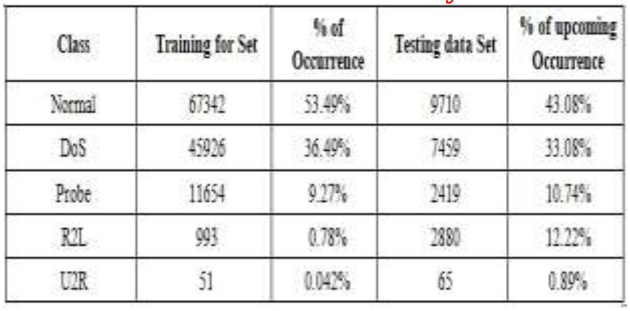
Dataset Used:

In our research, we will employ the NSL-KDD dataset, which is an enhanced variant of the KDD Cup dataset. In 1999, the KDD Cup dataset was developed for the International Knowledge Discovery and Data Mining tool competition. Its objective was to accumulate instances of network traffic.

The primary objective of the competition was to develop a prediction model that could distinguish between secure and hazardous data packets. A total of 43 characteristics are present in each instance of the NSL-KDD dataset. Out of these, 41 attributes provide information about the traffic data that is being transmitted, and the final two attributes indicate whether the data is indicative of a normal traffic flow or an attack. The author of this work describes these characteristics in great detail.

The NSL-KDD dataset contains a variety of files that are used for testing and training. Normal, DoS, Probe, R2L, and U2R are the five distinct categories into which the 125,973 events in the training set are divided. Similarly, the test set comprises 22,544 examples that are categorized into the same five categories. Table 1 provides a comprehensive examination of the instance numbers in each class. The distribution of instances in the training set and the testing set is illustrated in Figures 1 and 2.

Table 1: No. of Instances in Each Class



Class	Training for Set	% of Occurrence	Testing data Set	% of upcoming Occurrence
Normal	67342	53.49%	9710	43.08%
DoS	45926	36.49%	7459	33.08%
Probe	11654	9.27%	2419	10.74%
R2L	993	0.78%	2880	12.22%
U2R	51	0.042%	65	0.89%

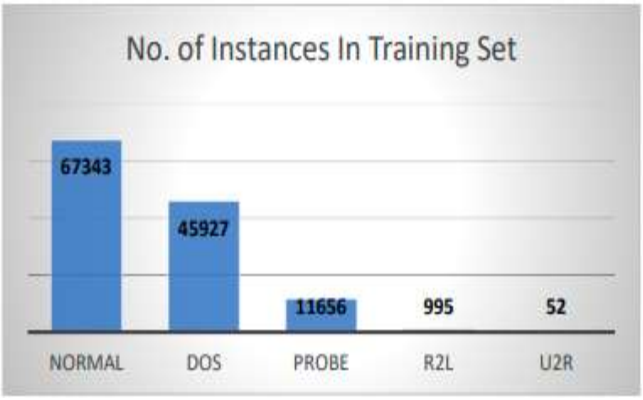


Fig 1: No. of Instances in Training Set

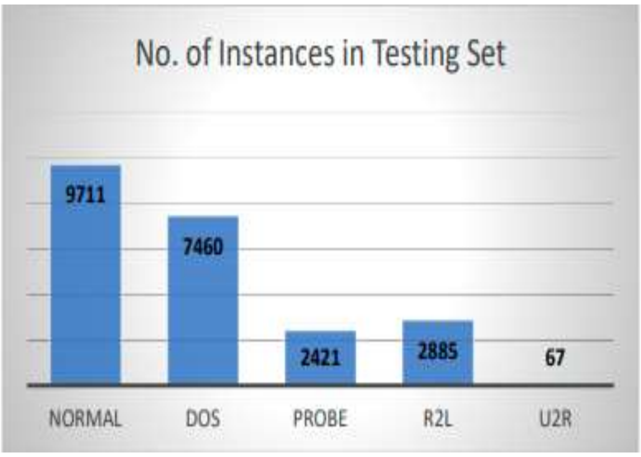


Fig 2: The number of instances in the test set

Within the NSL-KDD using dataset, there are 4 distinct classes for attack:

i. DoS (Denial of Service): This attack prevents the network's intended users from accessing it by sending an excessive number of requests simultaneously. This is exemplified by the SYN Flooding event.

ii. Probe or Surveillance: This method enables the perpetrator to access the data on a remote computer and exploit it for their own malicious purposes. One example of this form of attack is port scanning.

iii. U2R (User to Root): A right to employ a U2R attack in order to attempt to obtain root privileges, thereby exposing the machine to attack. For an illustration of this, examine a buffer overflow attack.

iv. R2L (Remote to Local): A remote-to-local (R2L) attack is initiated by an individual who attempts to exploit vulnerabilities in the victim's system to access it from a distance. An example of this form of hack is the attempt to guess a password.

3. PROPOSED METHODOLOGY

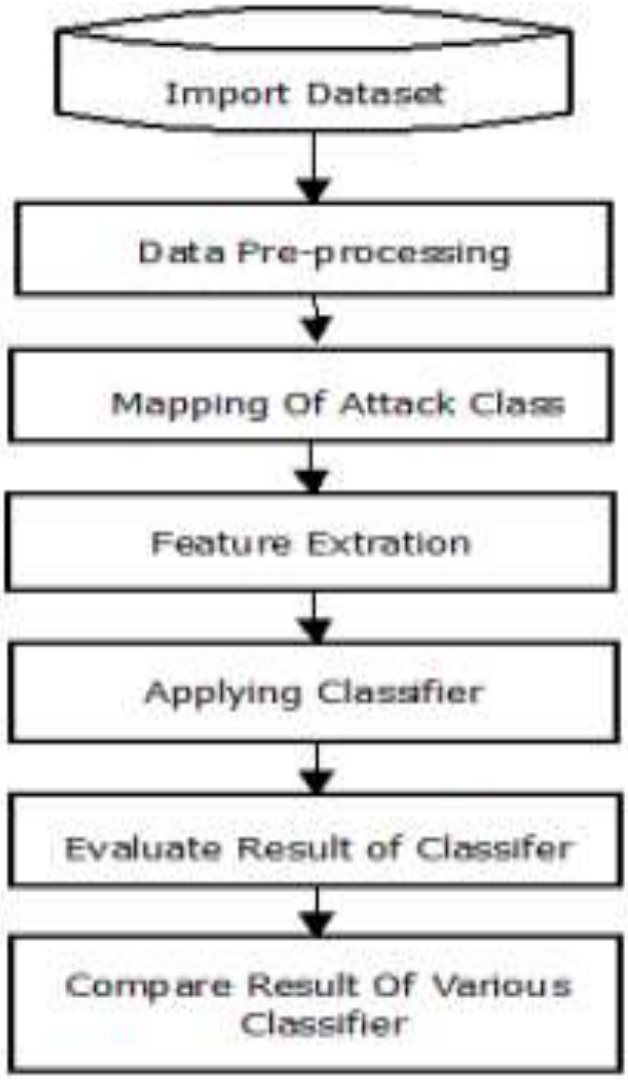


Fig 3: Methodology Used

The methodology adopted for the research consists of the following steps:

Data Pre-Processing: A dataset is selected from the NSL_KDD dataset repository, and features that are not numbers are converted to numbers using preprocessing techniques.

Mapping: Different attack classifications are assigned to different types of strikes.

Feature Selection: At order to eliminate skewed training, dimensionality reduction techniques like as random sampling are employed at this step.

Applying Classifier: A number of machine learning methods are applied to group the data.

Evaluating Performance Metrics: The accuracy, precision, and recall of the classifiers are evaluated based on several variables. A schematic illustrating the actions taken is shown in Figure 3.

Performance Evaluation and Experimental Analysis:

We examined the success metrics listed in Table 2, which is displayed below.

Table 2: Metrics for Performances

Metrics		Actual Class	
		A	NotA
Predicted Class	A	TRP	FLP
	Not A	FLN	TRN

The used performance metrics are as follows:

$$Accuracy = \frac{TRP + TRN}{TRP + FLP + FLN + TRN}$$

$$Recall = \frac{TRP}{TRP + FRN}$$

$$Precision = \frac{TRP}{TRP + FLP}$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

The experiment's findings are as follows: For our investigation, we selected 20,000 cases from the training and test sets of the NSL-KDD dataset. The frequency percentage of data from the assault class or normal data is depicted in Figure 4. A comprehensive list of the presentation metrics for all seven classes utilized in the experiment is also provided in Table 3. The metrics in question are precision, recall, and accuracy.

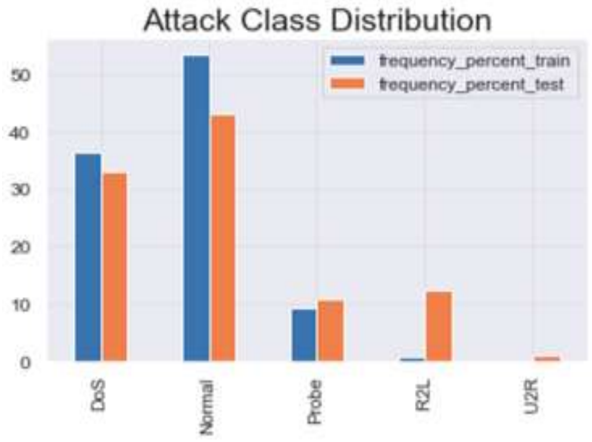
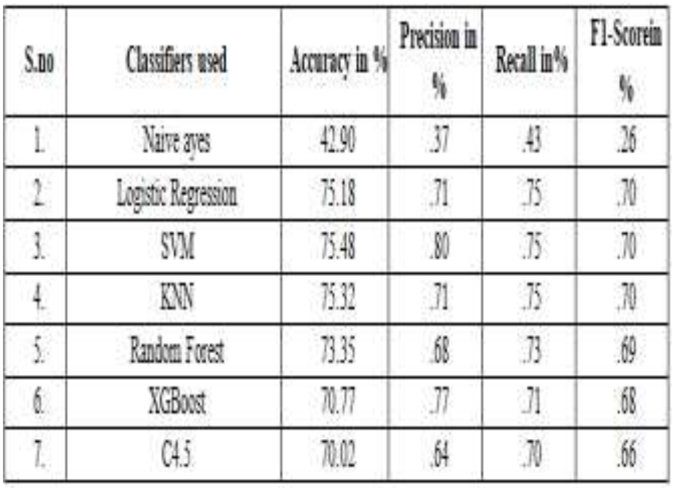


Fig 4: Attack Class Distribution

Table 4: Provides a concise summary of all classifiers' performance metrics.



S.no	Classifiers used	Accuracy in %	Precision in %	Recall in%	F1-Scorein %
1.	Naive ayes	42.90	.37	.43	.26
2.	Logistic Regression	75.18	.71	.75	.70
3.	SVM	75.48	.80	.75	.70
4.	KNN	75.32	.71	.75	.70
5.	Random Forest	73.35	.68	.73	.69
6.	XGBoost	70.77	.77	.71	.68
7.	C4.5	70.02	.64	.70	.66

As illustrated in Figure 5, the SVM algorithm obtained the highest degree of accuracy. Figure 6 demonstrated that SVM outperformed other models in terms of accuracy. Furthermore, Figure 7 illustrates that the optimal recall rates were achieved by the SVM, KNN, and Logistic classifiers.

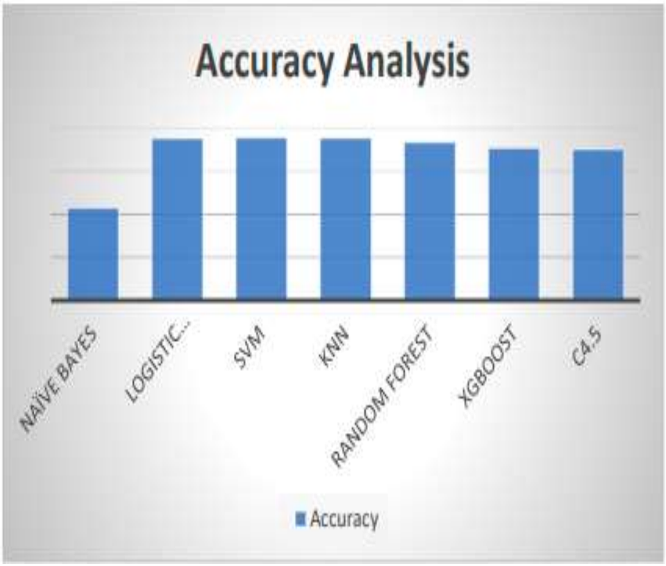


Fig 5: Accuracy Analysis

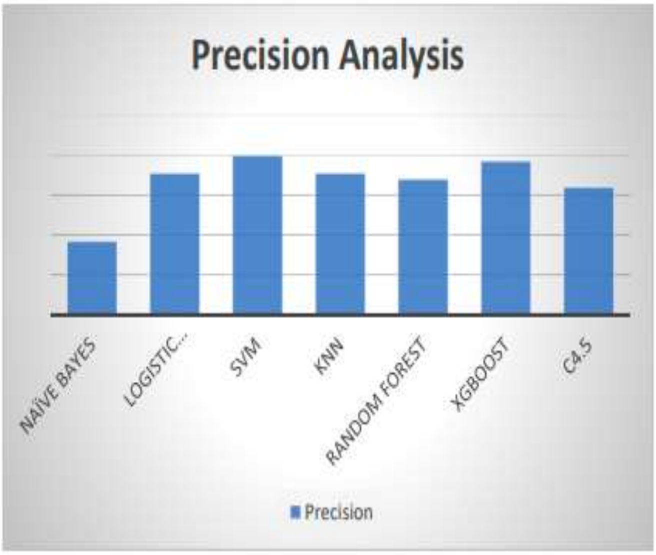


Fig 6: Precision Analysis

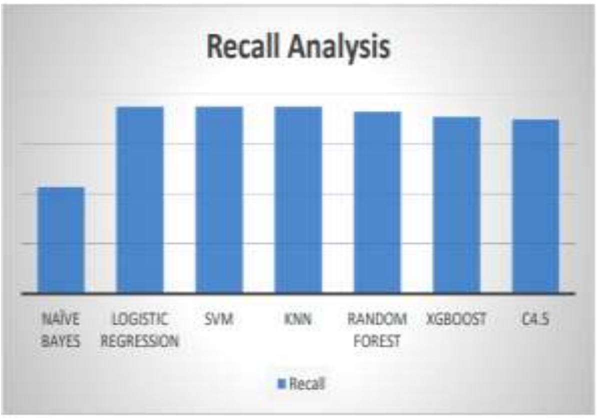


Fig 7: Recall Analysis

4. CONCLUSION AND FUTURE WORK

In order to conduct a comparison, we implemented Naïve Bayes and Logistic Regression using the scikit module in Python. Regression, XG Boost, Random Forest, KNN, C4.5, and SVM are all examples of classifiers. The SVM method achieved the highest levels of precision, recall, and accuracy on the NSL KDD dataset after preprocessing the data. The KNN predictor was not as accurate as the SVM, despite the fact that they were executed at significantly distinct times. The results of our experiments indicate that SVM outperforms other classifiers in terms of precision, memory, and accuracy. We will enhance the functionality of machine learning algorithms in our new approach by employing them on real-time data. Additionally, we are interested in identifying assaults that are not included in the dataset.

REFERENCES

1. Dewa, Leandros Maglaras (2016) “Data Mining and Intrusion Detection Systems”, International Journal of Advanced Computer Science and Applications, Vol 7 No 1,pp61-71.
2. L. Dhanabal, and S. P. Shantharajah (2015) “A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms”,International Journal of Advanced Research in Computer and Communication Engineering, Vol 4, Issue 6,pp.
3. Muhammad Shafiq, Xiangzhan Yu, Asif Ali Laghari, Lu Yao, N abin Kumar Karn, F oudil Abdessamia, “Network Traffic Classification Techniques and Comparative Analysis Using
4. Machine Learning Algorithms”, 2016 2nd IEEE International Conference on Computer and Communications, vol. 8, pp. 2451-2455, 2016.
5. JaiswalRupeshChandrakant, Lokhande Shashikant. D., “Machine Learning Based Internet Traffic Recognition with Statistical Approach”, 2013 Annual IEEE India Conference (INDICON), vol. 7, pp. 121-126, 2013.
6. RiyadAlshammari, A. NurZincir-Heywood, “Identification of KDD encrypted traffic using a machine learning approach”, Journal of King Saud University – Computer and Information Sciences, vol. 27, pp. 7792, 2015.
7. Alberto Dainotti, Antonio Pescapé, Kimberly C. Claffy,” Issues and Future Directions in Traffic Classification”, IEEE Network January/February 2012.
8. T.Nguyenand G.Armitage,“A survey of techniques for Internet traffic classification using machine learning”, IEEE Communications Surveys &Tutorials,Vol.10,No.4,fourthquarter2008,pp 56-76.
9. atihErtam, Ilhan FiratKilinçer, Orhan Yaman,”Intrusion Detection in Computer Networks via Machine Learning Algorithms”, International Artificial Intelligence and Data Processing Symposium(IDAP),2017,pp 1-4
10. Jamal H. Assi, Ahmed T. Sadiq, ”NSL-KDD dataset Classification Using Five Classification Methods and Three Feature Selection Strategies”, Journal of Advanced Computer Science and Technology Research, Vol.7 No.1, March 2017, 15-28.
11. Muhammad Shafiq, Xiangzhan Yu, Asif Ali Laghari, Lu Yao, N abin Kumar Karn, FoudilAbdessamia, “Network Traffic Classification Techniques and Comparative Analysis Using Machine Learning Algorithms”, 2nd IEEE International Conference on Computer and Communications,2016,pp2451-2455
12. Chaturvedi, Pooja, A. K. Daniel, and Vipul Narayan. "A Novel Heuristic for Maximizing Lifetime of Target Coverage in Wireless Sensor Networks." Advanced Wireless Communication and Sensor Networks. Chapman and Hall/CRC 227-242.
13. Kumar, Vimal, and Rakesh Kumar. "A cooperative black hole node detection and mitigation approach for MANETs." In Innovative Security Solutions for Information Technology and Communications: 8th International Conference, SECITC 2015, Bucharest, Romania, June 11-12, 2015. Revised Selected Papers 8, pp. 171-183. Springer International Publishing, 2015.
14. Kumar, V., Shankar, M., Tripathi, A.M., Yadav, V., Rai, A.K., Khan, U. and Rahul, M., 2022. Prevention of Blackhole Attack in MANET using Certificateless Signature Scheme. Journal of Scientific & Industrial Research, 81(10), pp.1061-1072.
15. Kumar, V. and Kumar, R., 2015, April. Detection of phishing attack using visual cryptography in ad hoc network. In 2015 International Conference on Communications and Signal Processing (ICCSP) (pp. 1021-1025). IEEE.