IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Advancing Network Security: Implementing Machine Learning Algorithms For Intrusion Detection

[1]Madhuri Wavhal, [2]Rashmi Ghadge, [3]Swaraj Gadge, [4]Omkar Shahane, [5]Suvarna Potdukhe,
[1,2,3,4]Student, [5]Professor,
[1,2,3,4,5]Information Technology,
[1,2,3,4,5]Rmd Sinhgad Technical Institute Campus, Warje Pune, India

Abstract: Intrusion Detection Systems (IDS) are essential for protecting against a wide range of cyberthreats in modern network security. In order to improve the effectiveness of intrusion detection systems, this paper provides an implementation methodology that focuses on four different machine learning algorithms: Support Vector Machines (SVM), Random Forest, Decision Trees, and Naive Bayes. Every algorithm is carefully customized and included into the IDS architecture to meet certain needs presented by contemporary network environments. Each algorithm's effectiveness is evaluated in terms of computing efficiency, false positive rates, and detection accuracy through extensive testing and analysis. The usefulness of the suggested framework in detecting different kinds of intrusions while reducing false alarms is verified using real-world datasets. In addition, issues of scalability, flexibility, and deployment feasibility are covered in order to promote smooth integration across various network infrastructures. For researchers, practitioners, and organizations looking to implement reliable and adaptable intrusion detection systems using cutting-edge machine learning methods, this paper is an invaluable resource.

***Index Terms* - machine learning; deep learning; support vector machine; intrusion detection system; cyber security**

I. INTRODUCTION

Protecting sensitive data and vital infrastructure from cyber assaults has grown more difficult in today's digitally connected society. At the forefront of defense are intrusion detection systems, or IDS, which watch over network traffic continually in order to identify and stop hostile activity and unauthorized access. The dynamic strategies employed by cybercriminals are often too much for conventional intrusion detection systems to keep up with. However, intrusion detection has undergone a radical change thanks to the incorporation of cutting-edge technologies like machine learning (ML) and deep learning (DL).

Machine Learning and Deep Learning techniques offer the promise of enhanced accuracy and efficiency in identifying both known and unknown threats within vast streams of network data. Unlike traditional rule-based approaches, ML and DL empower IDS to adapt and learn from data patterns, enabling them to detect anomalies and suspicious behaviors that might evade conventional detection methods. Several algorithms, including Support Vector Machines (SVM), Decision Trees (DT), Naive Bayes (NB), and Random Forests (RF), have shown promise in this domain, each with its unique strengths in capturing complex patterns and distinguishing between normal and malicious activities.

This study investigates the interrelationships of Machine Learning, Deep Learning, and Intrusion Detection Systems by examining the fundamental ideas, approaches, and practical uses of SVM, DT, NB, and RF algorithms. We look at how these algorithms might enhance the detection capabilities of IDS in a variety of business contexts and organizational settings by analyzing network traffic, system logs, and other data sources to find possible threats.

We also go over the opportunities and difficulties of using these algorithms to intrusion detection, such as adversarial attacks, model interpretability, and data scarcity. Through an awareness of the advantages and disadvantages of SVM, DT, NB, and RF, enterprises may plan and implement reliable intrusion detection systems (IDS) that are customized to meet their unique cybersecurity requirements.

II. Literature Survey

2.1. Paper Title: "Efficient Network Intrusion Detection and Classification System"

Authors: Iftikhar Ahmad, Qazi Emad Ul Haq

Abstract: The proliferation of networked systems and the ever-evolving cyber threat landscape underscore the pressing need for robust intrusion detection and classification systems. This paper introduces an innovative approach to bolster the security of modern network environments. Our system utilizes advanced machine learning techniques and a streamlined architecture for real-time detection and classification of network intrusions. This real-time operation ensures prompt threat identification and response to mitigate potential damage and safeguard critical assets.

2.2 Paper Title: "Network Traffic Analysis and Intrusion Detection with Packet Sniffer"

Authors: Mohammed Abdul Qadeer, Mohammad Zahid

Abstract: In today's interconnected digital world, network security is paramount. Real-time network traffic monitoring and intrusion detection are pivotal for safeguarding assets and maintaining network integrity. This paper presents a novel approach that leverages a high-performance packet sniffer for efficient network packet capture and analysis. We employ advanced traffic analysis to identify normal patterns, perform protocol analysis, and detect deviations from established baselines. Our intrusion detection system classifies threats, such as denial of service (DoS) attacks, intrusion attempts, and malware propagation in real-time, enabling rapid incident response.

2.3 Paper Title: "Multiclass Classification Baselines for Anomaly-based Network Intrusion Detection Systems"

Authors: Ajay Shah, Sophine Clachar, Manfred Minimair, Davis Cook

Abstract: Network Intrusion Detection Systems (NIDS) are pivotal in safeguarding computer networks. Anomaly-based NIDS, which rely on identifying deviations from normal network behavior, effectively detect known and unknown threats. This paper addresses the challenge of developing comprehensive multiclass classification baselines for anomaly-based NIDS. We propose a systematic approach that utilizes diverse datasets and machine learning techniques. By evaluating various classifiers, including traditional algorithms and deep learning models, we provide insights into their suitability for NIDS.

2.4 Paper Title: Comparing the Performance of Adaptive Boosted Classifiers in Anomaly based Intrusion Detection System for Networks

Abstract:

The computer network is used by billions of people worldwide for variety of purposes. This has made the security increasingly important in networks. It is essential to use Intrusion Detection Systems (IDS) and devices whose main function is to detect anomalies in networks. Mostly all the intrusion detection approaches focuses on the issues of boosting techniques since results are inaccurate and results in lengthy detection process. The major pitfall in network based intrusion detection is the wide-ranging volume of data gathered from the network. In this paper, we put forward a hybrid anomaly based intrusion detection system which uses Classification and Boosting technique. The Paper is organized in such a way it compares the performance three different Classifiers along with boosting. Boosting process maximizes classification accuracy. Results of proposed scheme will analyzed over different datasets like Intrusion Detection Kaggle Dataset and NSL KDD. Out of vast analysis it is found Random tree provides best average Accuracy rate of around 99.98%, Detection rate of 98.79% and a minimum False Alarm rate.

III. Methodology:

First, the IDS's goals and scope are outlined, along with the kinds of threats it seeks to identify and the network or systems it will keep an eye on. After that, data is collected from pertinent data sources, including system and network traffic logs. The obtained data is then subjected to preprocessing, which includes operations like noise reduction and normalization, to clean it up and get it ready for analysis. The next step is to pick or extract features in order to determine which attributes are most important for differentiating between harmful and legitimate activity. Next, utilizing the preprocessed data, appropriate deep learning and machine learning models are chosen, trained, and assessed. To improve performance, model optimization can be carried out, and ensemble approaches can be taken into consideration for improved detection capabilities. The IDS is deployed in the production environment and its performance is tracked after it has been trained. To remain successful over time and adjust to changing threats, regular maintenance and upgrades are necessary. Organizations can create reliable intrusion detection systems (IDS) to efficiently defend against cyberattacks by using this methodology.

3.1 SVM:

The most sophisticated and successful support vector machine classification methodology is the kernel function method. Using a kernel function, the "dimension disaster" issue with conventional classification techniques can be resolved successfully.

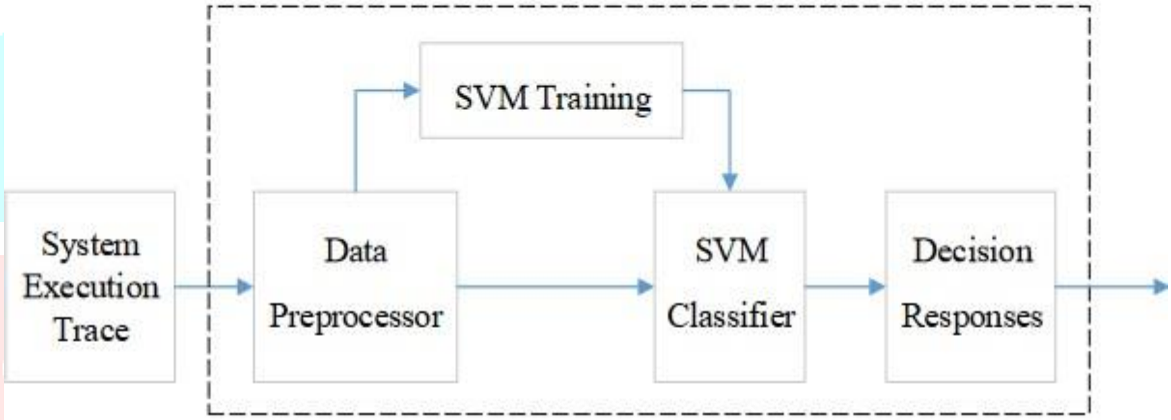

Figure 1. structure of intrusion detection system based on svm

Generally, a nonlinear mapping is used to translate the original input sample space to a high-dimensional feature space, and a kernel function meeting Mercer's condition is chosen. Support vector machines' linearly separable method can be applied in this manner to tackle the nonlinear non-separable problem. The following are examples of frequently used kernel functions:

(1) Polynomial Function
$K(x , x) = [(xi\ x)+1]^{q}$
(2) Gaussian Kernel Function (Radial Basis Function, RBF)
Radial Basis Function (RBF) Kernel:
$K(xi,xj)=\exp(-\gamma\| xi-xj \|^{2})$
(3) Sigmoid Function
$K(xi , x) = \tanh[v(xi \cdot x)+ c]$

The experiment used KDDCUP99 as the intrusion detection dataset in the publication. Every TCP/IP connection in the data collection has a corresponding record. There are forty-one qualities (attributes) that define each record. There are 22 attacks in the original data set, which may be categorized into four primary categories: R2L (remote unauthorized access), U2R (illegal access to the local super user), DOS (denial of service attacks), and probing (scanning and detection). All attack types are categorized as abnormal mode in the experiment, therefore intrusion detection just needs to determine if an attack is in normal mode or abnormal mode. The KDDCUP99 data set observation indicates that the data set is a heterogeneous data set. .

The 41 features (attributes) of each record have both textual and numerical values, some of which vary widely, while others have only two values, 0 and 1. For such a heterogeneous problem, data normalization is needed.

3.2 **Decision Tree:**

Decision trees are an effective technique for spotting malicious activity in a system or network. These trees are built using a dataset made up of different attributes that have been taken out of system logs or network traffic. Recursively dividing the dataset into subsets of instances with comparable properties is the basic idea underlying decision trees in IDS. This procedure goes on until a predetermined stopping criterion is satisfied, which usually happens when splitting the data further doesn't significantly increase classification accuracy or when all instances inside a subset belong to the same class.

To divide the dataset, the decision tree algorithm chooses the optimal characteristic at each node. The selection criterion seeks to optimize the homogeneity of instances within resulting subgroups and is frequently quantified using metrics such as information gain or Gini impurity. Mathematically, information gain can be expressed as:

$$IG(D,A)=H(D)-H(D \mid A)$$

where

IG(D,A) represents the information gain achieved by splitting dataset

D based on feature A.

H(D) denotes the entropy of dataset D, which measures its impurity or disorder.

H(D∣A) represents the conditional entropy of D given feature A, indicating the remaining uncertainty after considering feature A.

The dataset is further divided by the decision tree in a cyclical manner until it reaches leaf nodes, which stand for the final classification results. These results often correlate to either benign or malicious activity in the context of intrusion detection.

3.3 Random Forest:

Random forest modeling algorithm for networks IDS NSL-KDD dataset as input Classification of various attack types as an output

Step 1: open the dataset.

Step 2: Utilize the pre-processing method Separation

Step 3: Divide the dataset into four groups in step three.

Step 4: Divide the dataset into test and training sets.

Step 5: Use the feature subset selection measure to choose the optimal collection of features. Uncertainty symmetry (SU)

Information gain is compensated by symmetric uncertainty.

$$IG\ (X/Y)/H(X)H(Y)] = 2[SU(X, Y\ )]$$

Step 6: Random Forest receives the data set for training.

Step 7: Random Forest is then used to classify the test data set.

Step 8: Determine the Mathew correlation coefficient, accuracy, detection rate, and false alarm rate.

In ARFF format, we downloaded the NSL-KDD dataset for our experimental study. For the experiment, we used the following preprocessing methods.

1) Fill in any missing values: To replace every missing feature value in the NSL-KDD dataset, we employed the replace missing values filter.

The mean and mode from the training set of data are used by this filter to replace any missing values.

2) Discretization: Using unsupervised 10 bin discretization, discretization filters were used to discretize numerical attributes.

3.4 Naïve Bayes:

This model computes the probability of a final outcome based on multiple connected evidence variables. It is a simplified version of the Bayesian probability model. It makes the assumption that these variables are independent of one another, which means that the likelihood of one feature occurring has no bearing on the likelihood of another. This model can be used in intrusion detection scenarios where evidence variables are things like weather or seismic activity, such as alarm monitoring for theft detection.

Mathematical Equation: $$P(C|F_1, F_2,\ldots, F_n) = \frac{P(C).P(F_1|C).P(F_2|C)\ldots..P(F_n|C)}{P(F_1)\cdot P(F_2)\cdot\ldots\cdot P(Fn)}$$

The class variable of importance in this structure is the likelihood of theft, with other attributes acting as evidence of likelihood. Under the strong independence assumption, the model makes 2^n! independent assertions given n attributes. Nevertheless, naïve Bayes classifiers often yield precise outcomes.

Research on factors impacting classifier performance has revealed three main sources of error: noise, bias, and variation in training data. Using high-quality training samples is the only technique to minimize noise in training data. While bias originates from excessively large groupings, too small groupings are the source of variance in data. It is crucial to find a balance between these variables in intrusion detection scenarios in order to enhance classifier accuracy and minimize errors.

IV. System Architecture:

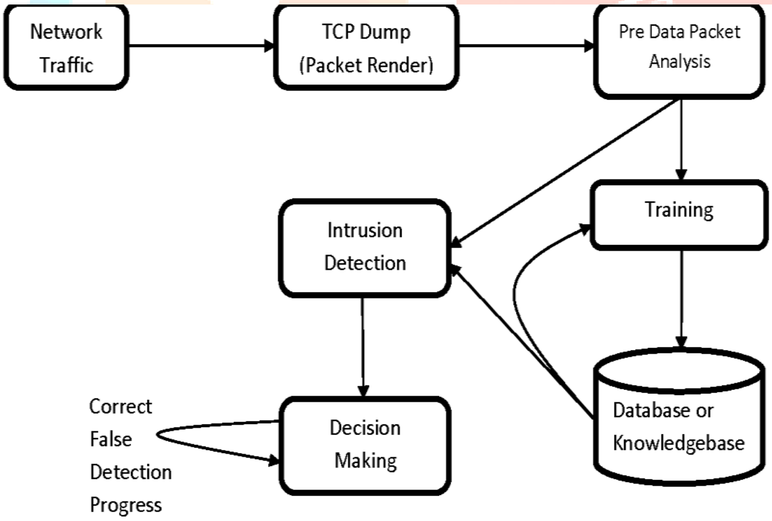

Fig 2: system architecture of ids

V. Result

Serial No.	Classification algorithm	Accuracy Obtained
1	SVM	76.410
2	RF	92.820
3	DT	92.820
4	NB	41.538

Table 1: accuracy comparison table

VI. Conclusion:

Cybercriminals employ social networking tactics and state-of-the-art technology to target computer users. There are hackers that are more motivated and skilled than others. Cybercriminals have proven that they are capable of hiding their identities, communications, and illegal income in addition to using strong infrastructure. As a result, using advanced IDSs to safeguard systems that can identify modern threats is becoming more and more crucial. A full grasp of the benefits and drawbacks of modern IDS research is essential for developing or creating such IDS systems. Additionally, we included a detailed review of the approaches, kinds, or technologies used in intrusion detection systems, as well as the benefits and drawbacks of each.

VII. References

1. William Stallings, “Cryptography and Network Security”, 5th edition, Pearson Education.
2. Simon Hansman. “A Taxonomy of Network and Computer Attack Methodologies” 2003, url: https://www.cosc.canterbury.ac.nz/research/reports/HonsReps/2003/hon s_030.pdf (visited on 11/09/2019).
3. Abhishek Pharate , Harsha Bhat , Vaibhav Shilimkar “Classification of Intrusion Detection System” IJCS Volume 118 – No. 7, May 2015.
4. 4. SNORT. (2017). Snort 2.9.7.6. [Online]. Available: https://www.snort.org/
5. OISF. (2018). Suricata 4.0.4. [Online]. Available: https://suricataids.org/about/
6. T. Sree Kala, Dr .A. Christy “A Survey and Analysis of Machine Learning Algorithms for Intrusion Detection System” Jour of Adv Research in Dynamical & Control Systems, 04-Special Issue, June 2017.
7. https://www.guru99.com/machine-learning-tutorial.html.
8. Thuy T.T. Nguyen and Grenville Armitage “A Survey of Techniques for Internet Traffic Classification using Machine Learning” IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 10, NO. 4, FOURTH QUARTER 2008.
9. Shilpa Bahl and Dr. Deepak Dahiya “Features Contribution for Detecting Attacks of an Intrusion Detection System” Global Journal of Pure and Applied Mathematics. ISSN 0973-1768 Volume 13, Number 9 (2017), pp. 5635-5653.
10. https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feat ure-selection-methods-with-an-example-or-how-to-select-the-right-vari ables/Sumaiya Thaseen Ikram , Aswani Kumar Cherukuri “Intrusion detection model using fusion of chi-square feature selection and multi class SVM” Computer and Information Sciences (2017) 29, 462–472.