Network Intrusion Detection System

|  |  |  |
| --- | --- | --- |
| Sheetal PhatangareDepartment of Computer Engineering Vishwakarma Institute of Technology, Pune, Maharashtra, Indiasheetal.phatangare@vit.edu | Shrinivas JadhavDepartment of Computer Engineering Vishwakarma Institute of Technology, Pune, Maharashtra, IndiaShrinivas.jadhav21@vit.edu | Sandesh KawaneDepartment of Computer Engineering Vishwakarma Institute of Technology, Pune, Maharashtra, IndiaSandesh.kawane21@vit.edu |
| Prajwal HolkarDepartment of Computer Engineering Vishwakarma Institute of Technology, Pune, Maharashtra, IndiaPrajwal.holkar21@vit.edu | Pratiksha GaikwadDepartment of Computer Engineering Vishwakarma Institute of Technology, Pune, Maharashtra, IndiaPratiksha.gaikwad21@vit.edu |  |

***Abstract:- The growing dependence on the internet has increased the vulnerability of businesses to cyber threats. Network Intrusion Detection Systems (NIDS) are essential tools for safeguarding sensitive data and network infrastructure by identifying and flagging suspicious activities that may indicate malicious intent. This paper introduces a novel NIDS that leverages a combination of machine learning algorithms to enhance anomaly detection and classification. The system employs Support Vector Machines, K-Nearest Neighbors, Random Forest, and Logistic Regression to effectively differentiate between legitimate network traffic and a diverse range of attack types. This multifaceted approach capitalizes on the strengths of each algorithm: SVMs for robustly classifying high-dimensional data, KNN for efficient multi-class classification, Random Forests for handling complex datasets and mitigating overfitting, and Logistic Regression for interpretability and scalability. The NIDS utilizes the NSL-KDD dataset, a widely recognized benchmark for NIDS evaluation, to train and assess model performance.***

***Keywords- Network Intrusion Detection System (NIDS), Machine Learning, Deep Learning, Long Short-Term Memory (LSTM), K-Nearest Neighbors (KNN), NSL-KDD Dataset, MERN Stack, ReactJS, NodeJS, MongoDB, Google OAuth 2.0***

1. INTRODUCTION

Network security traditionally revolved around building a fortified "protective shield" around systems. This involved user identification and authentication (I&A) to control access and encryption to safeguard data confidentiality, preventing unauthorized disclosure of information. Additionally, measures were implemented to ensure data integrity, guaranteeing the accuracy and trustworthiness of data during transmission. Security protocols also aimed to protect against Denial-of-Service (DoS) attacks, where attackers attempt to overwhelm a system with requests, rendering it unavailable to legitimate users. Mandatory access control techniques, sometimes leveraging cryptography, were employed to restrict unauthorized access to sensitive data and functionalities.

However, this prevention-based approach has inherent limitations:

* **Imperfect Systems:** Even the most meticulously designed systems with numerous components might harbor vulnerabilities. Human error in configuration or outdated equipment can further compromise security.
* **Legacy Infrastructure:** Replacing existing infrastructure with entirely secure systems is often impractical due to immense investment costs in the current setup.
* **User Constraints:** Prevention-based methods can restrict user activities, hindering the open and dynamic environment that fosters productivity.
* **Limited Defense:** Encryption can't safeguard against lost or stolen keys or cracked passwords. Additionally, insider threats, where authorized users misuse privileges, remain a challenge.

The ever-present threat of cyberattacks necessitates robust security measures for businesses operating in today's internet-driven world. Network Intrusion Detection Systems (NIDS) play a crucial role in safeguarding sensitive data and network infrastructure. They achieve this by continuously monitoring network traffic and identifying anomalies that may indicate malicious activity, such as unauthorized access attempts or data exfiltration efforts. This paper presents the development of a novel NIDS that leverages advanced Machine Learning (ML) and Deep Learning (DL) techniques to achieve enhanced attack detection and classification capabilities. By employing these techniques, the NIDS aims to provide businesses with a more comprehensive and effective defense against cyber threats.

1. RELATED WORK

In the pursuit of real-time network intrusion detection, Dong et al. [1] propose a system that leverages big data technologies and deep learning techniques. Their system tackles the challenge of efficiently processing massive volumes of log data for intrusion detection.The authors highlight the importance of real-time detection capabilities in today's security landscape. To address this, they incorporate big data technologies like Flume and Flink for efficient data acquisition and processing.

In the paper[2] "Detecting Intrusions and Attacks in the Network Traffic using Anomaly based Techniques" by Vinod Kumar, Vinay Choudhary, and Vivek Sahrawat, the authors outline the research design and chosen methods for data collection, including surveys and interviews. They describe how the collected data will be analyzed, ensuring the reader that the research was conducted fairly and thoroughly. This detailed explanation allows readers to understand and potentially replicate the study..

In the paper "Anomaly-Based Intrusion Detection System for Ad hoc Networks" by Abdelaziz Amara Korba, Mehdi Nafaa, and Yacine Ghamri-Doudane, the authors propose a system that detects wormhole and rushing attacks in ad-hoc networks using a statistical approach[3] within a hierarchical network structure. Cluster heads collect data on node selection rates during route discovery, and nodes with unusually high selection rates are flagged as potentially malicious. The system also considers network mobility and density, aiming to detect attacks while improving network load balancing.

In the research "Network Intrusion Detection System Model Based on Data Mining"[4] by Yanjie Zhao, the author discusses the methodology as the roadmap for conducting research. The paper outlines specific methods used to gather and analyze data, which ensures that the research is transparent and repeatable. This can involve surveys, interviews, experiments, or other techniques, depending on the research question. Understanding the methodology allows readers to assess the credibility of the research and its findings.

In the paper "A Practical Network-based Intrusion Detection and Prevention System" by N. Wattanapongsakorn, S. Srakaew,[5] and C. Charnsripinyo, the authors propose a method to build a system that automatically protects computer networks from attacks. The system extracts key information from network traffic and employs various machine learning techniques to identify suspicious activity. Based on the results, it blocks malicious traffic. This approach was tested in a simulated environment and demonstrated promising results in detecting and preventing various cyberattacks.

In the paper[6] "Analyzing the Performance of Machine Learning Algorithms in Anomaly Network Intrusion Detection Systems" by Pascal Maniriho and Tohari Ahmad, the researchers compared several machine learning algorithms to determine which ones performed best at identifying network attacks. Using a dataset of known attacks and non-attacks, they tested different algorithms with and without feature selection techniques. Random Forest was found to be the most effective, but the researchers recommend exploring other algorithms, tuning the settings of those they tested, and considering newer techniques like deep learning. They also emphasize the importance of being aware of potential biases in the data, such as the underrepresentation of certain attack types, and suggest methods to mitigate these biases.

Due to deep learning's[7] capacity to handle massive datasets with excellent accuracy and short training times, intrusion detection systems (IDSs) have been using it more and more in recent years. The ability to detect novel and complex attacks is a shortcoming of traditional intrusion detection systems (IDS), such as signature-based and anomaly-based techniques. Dynamic, real-time intrusion detection is made possible by deep learning approaches, which provide a strong alternative with their hierarchical feature extraction capabilities. Karatas et al.'s survey paper gives a summary of the several deep learning algorithms used in IDSs, compares how effective they are, and talks about the datasets that are frequently utilised in this field. The authors stress the benefits of deep learning for enhancing the precision and effectiveness of intrusion detection systems, especially when managing intricate and dynamic cyberthreats. For researchers wishing to apply deep learning methods to improve IDS capabilities, this work provides an extensive reference.

Machine learning[8] has gained popularity in network intrusion detection in recent years because of its capacity for real-time analysis and flexibility. Ahmed et al. (2016) highlighted the drawbacks of conventional signature-based systems by classifying anomaly detection techniques into statistical, classification, clustering, and information theory methods. Studies on clustering algorithms by Syarif et al. (2017) and K-Means optimisations by Eslamnezhad et al. (2018) showed differing degrees of efficacy in anomaly detection. Promising outcomes have also been observed with Radial Basis Function (RBF) networks and Support Vector Machines (SVMs); Bo and Yan (2020) optimised RBF networks for real-time detection, while Zhu and Liao (2019) enhanced SVM efficiency. Moreover, Lakhina et al. (2017)'s discussion of the application of entropy metrics for anomaly detection has been crucial in raising detection accuracy. The necessity for ongoing research into cutting-edge machine learning methods and reference datasets like Kyoto 2006+ is highlighted by this changing landscape.

The paper "Machine Learning Approach to IDS: A Comprehensive Review" [9]examines the development and efficacy of machine learning-enhanced intrusion detection systems (IDS). Since its invention by Dorothy E. Denning in the 1980s, intrusion detection systems (IDS) have progressed from signature-based systems, which identify known attacks, to anomaly-based systems, which spot departures from the norm but have a high false alarm rate. By strengthening anomaly detection capabilities, machine learning—including supervised, semi-supervised, and unsupervised learning—has greatly enhanced IDS. Techniques for reducing data, like feature extraction and selection, are essential for controlling dataset complexity and enhancing classification accuracy. For assessing the efficacy of IDS, performance measurements such as ROC curves and confusion matrices are crucial. Though problems like false positives and negatives still exist, recent developments have focused on hybrid and ensemble approaches, which combine many classifiers to increase detection accuracy and decrease errors.

A thorough analysis of the use of machine learning techniques in network intrusion detection[10] systems (NIDS) can be found in the publication "Network Intrusion Detection Using Machine Learning Techniques" by Yasmeen S. Almutairi et al. It talks about the shortcomings of conventional intrusion detection systems (IDS), which mostly depended on rule-based and pattern recognition systems, emphasising their incapacity to defend against novel or complex threats. In order to enhance the detection of anomalies and attacks, the study centres on the application of multiple machine learning techniques, such as Support Vector Machines, J48, Random Forest, and Naïve Bayes. The paper evaluates these algorithms' performance in binary and multiclass classification scenarios using the NSL-KDD dataset. The findings demonstrate that by increasing detection accuracy and decreasing false positives, machine learning techniques provide notable advantages over conventional techniques, thereby tackling some of the most important issues in contemporary network security.

The paper [11]"Network Intrusion Detection Leveraging Machine Learning and Feature Selection" by Arshid Ali et al. provides a thorough analysis of how to improve intrusion detection systems (IDS) by combining feature selection approaches with machine learning (ML) techniques. Using methods like Correlation-based Feature Selection (CFS) and Classifier Subset Evaluation with Naïve Bayes (NB), the authors tackle the difficulties associated with handling high-dimensional data in IDS. Their work uses the CIC IDS-2017 dataset to assess two machine learning algorithms, Multilayer Perceptron (MLP) and Instance-Based Learning (IBK), in order to illustrate the efficacy of these techniques. The findings show that feature selection enhances the accuracy of the IBK classifier to remarkable levels—99.87% and 99.82% for various feature selection techniques—beyond MLP in both model-building time and accuracy. In order to simplify and improve IDS performance, the paper emphasises the significance of choosing pertinent features. It also offers insightful information about the effectiveness of machine learning algorithms in network security applications.

The increasing significance of[12] machine learning approaches in augmenting network security for Network-based Intrusion Detection Systems (NIDS) is highlighted by the literature. The inability of traditional IDS techniques to identify novel or unidentified attacks has prompted the use of machine learning techniques. According to research, machine learning can greatly increase detection rates by picking up on complex patterns that point to both known and unknown threats by studying vast datasets. Numerous research have demonstrated the promise of techniques like ensemble methods, decision trees, and neural networks. Neural networks are good at processing complex patterns in data, while decision trees provide interpretability. New developments also highlight hybrid models, which mix various machine learning techniques to maximise efficiency. Problems like the high false-positive rate and the requirement for adaptable systems that can change in response to new threats are the focus of continuing research.

The challenges of integrating machine learning[13] into network intrusion detection systems (NIDS) are examined in Sommer and Paxson's study, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection". Machine learning has shown promise in other areas, such as spam and product recommendations, but its use in network security anomaly detection has been restricted by issues like the closed-world assumption, high error costs, and the semantic gap between anomalies and actionable insights. The study makes the case that, despite its potential benefits, machine learning cannot effectively handle these particular difficulties unless it is thoroughly understood to include the unique operational context and constraints of network security.

A thorough analysis of several machine learning techniques[14] used with intrusion detection systems (IDS) can be found in the paper "Intrusion Detection System using Machine Learning Techniques: A Review". It illustrates the progression of intrusion detection systems (IDS) from conventional signature-based approaches, which depend on well-known attack patterns, to sophisticated machine learning techniques, which increase detection accuracy and address new threats. The overview includes ensemble approaches that integrate the output of numerous models to improve predicted accuracy, hybrid classifiers that mix various algorithms for greater performance, and single classifiers like SVM and decision trees. In contrasting these methods across several investigations, the research highlights the benefits of hybrid and ensemble strategies in lowering false positives and raising detection rates, while also pointing out the continuous difficulties in adjusting IDS to changing threats.

In order to increase network security, [15]this study provides an intrusion detection system (IDS) that uses machine learning to detect hostile activity more accurately. Using the NSL-KDD dataset, the study focusses on applying the Support Vector Machine (SVM) and Naïve Bayes methods to classify network traffic data and detect intrusions. The outcomes show that SVM performs better in terms of accuracy and less misclassification rates than Naïve Bayes. To maximise the effectiveness of the IDS, the suggested method incorporates crucial procedures such feature reduction, data pre-processing, and normalisation. The study shows that compared to Naïve Bayes, SVM obtains a greater accuracy rate, making it a more useful tool for intrusion detection.

1. FLOWCHART



Fig01:- Flow chart of system

This section describes the network intrusion detection system (NIDS) process through a flowchart (insert flowchart reference here). The flowchart details the following key components and their interactions:

Data Acquisition:

* **System Logs:** Security events and relevant information are captured from system logs.
* **Network Traffic Capture:** Network data packets are captured for analysis.

Data Preprocessing:

* The captured data undergoes preprocessing to prepare it for analysis. This may involve:
	+ Filtering out irrelevant data (e.g., internal network traffic).
	+ Formatting the data into a standard format suitable for analysis by the NIDS engine.

Data Analysis and Detection:

* **Database:** The preprocessed data is stored in a database for efficient retrieval and analysis.
* **NIDS Rule Base:** The NIDS engine leverages a set of pre-defined rules to identify suspicious activity within the network traffic. These rules may be based on known attack patterns, network traffic anomalies, or specific protocols.
* **Classification Generation:** The NIDS engine analyzes the data against the rule base. If a match is found, a classification is generated, indicating potential malicious activity.

Output and Response:

* **Console:** The classification results are displayed on the console, alerting security personnel of potential threats. This may include details about the detected anomaly and its severity.
* **Temporary Documents (Optional):** Depending on the system configuration, the NIDS may store the data in temporary documents for further investigation or historical analysis, particularly for unidentified events.
1. METHODOLOGY

Dataset pre-processing, classification and result evaluation are the vital phases in the proposed model. In proposed system each phase is essential and enhances important influence on its performance. To examine the function of SVM and Naïve Bayes classifiers are the essential steps of this work.

1. Data preprocessing

Dataset contains symbolic features; these features are unable to process by the classifier. Hence, pre-processing takes place. In this phase all non-numeric or symbolic features get removed or exchanged. Elimination or replacement of non-numeric or symbolic features is done in pre-processing phase. The overall process of pre-processing is essential, in which non-numeric or symbolic features are eliminated or replaced, as they do not perform any important participation in intrusion detection. Symbolic attributes like protocol, service and flag get changed or removed. Finally, the instances get labeled under four categories: Normal, DoS, Probe, and R2L.

1. Methodology
* Comparative analysis done between SVM and Naïve Bayes for classification of dataset, to analyze their accuracy and Misclassification Rate. At first raw dataset is taken and the class attribute contains 24 different types of attack which get labeled under 4 categories. They are normal, Dos, Probe, r2l.
* After, labeling Pre-processing is done to convert nominal attribute to binary attribute. In order to obtain improved performance of intrusion detection system, non-numeric features get removed.
* For randomization, the dataset is allowed to get processed in WEKA tool by incorporating the filter Randomize. Randomize filter randomly shuffles the order of instances passed through it by setting a random number generator, in which the seed value get reset. Collecting the first 19,000 instances for comparative analysis.
* In order to get different result and to improve the performance of the dataset, methodologies like CfsSubsetEval is done for feature reduction. The given dataset after preprocessing under goes feature reduction and normalization.
* CfsSubsetEval is one of the methods of attribute selection. It calculates the value of attributes by considering the individual predicting estimation of all features along with the degree of redundancy between them.
* About classification under SVM, it comes under supervised learning method, in which various types of data from different subjects get trained. In a given high dimensional space, Support Vector Machine creates hyperplane or multiple hyperplanes in a highdimensional space. SVM creates hyperplane or multiple hyperplanes.
* The hyperplane which optimally separates the given data into various classes with the major partition, consider as a best hyperplane. For evaluate the margins between hyperplanes, a non-linear classifier applies various kernel functions. Maximizing margins between hyperplanes is the main aim of these kernel functions like linear, polynomial, radial basis, and sigmoid.
* For the first 19,000 instances, classification of raw dataset using SVM, SVM under different Normalization techniques and SVM along with Feature Reduction is done for comparative analysis. Accuracy and Misclassification rate also noted.
* Same, process is done using Naïve Bayes. Bayesian classifiers are statistical classifiers. They are capable to forecast the probability that whether the given model fits to a particular class. It is based on Bayes’ theorem. It works on the hypothesis that, for a given class, the attribute value is independent to the values of the attributes. This theory is called class conditional independence.
* Naive Bayes classifier works as follows: Training set of samples get denoted by T, each with their class labels. There are k classes, X1,X2,X3,..,X(k-1),Xk. A = {a1,a2,...,an}, depicting n measured values of the n attributes, m1,m2,...,mn, whereas A depicting n-dimensional vector,. b) For a given sample A, the classifier will calculate A, which fits to the class having the maximum posteriori probability, conditioned.
1. DATASET DESCRIPTION

The dataset used in this work is the KDD Cup 1999 data [1], a benchmark dataset for network intrusion detection. It was originally used in the Third International Knowledge Discovery and Data Mining Tools Competition held in conjunction with KDD-99 [1]. The task of the competition was to build a network intrusion detection system capable of classifying network connections as either normal or belonging to a specific attack category.

The KDD Cup 1999 data consists of a collection of labeled network connection vectors. Each vector represents features extracted from a single network connection and is classified as either "normal" or an attack type. The attack types are further categorized into four main groups:

\* Denial-of-Service (DoS): Attacks that aim to disrupt service to legitimate users.

\* User to Root (U2R): Attacks that allow an attacker to gain unauthorized root access on a local machine.

\* Remote to Local (R2L): Attacks that allow an attacker to gain unauthorized access to a machine on a network from a remote location.

\* Probing: Attempts to gather information about a network or system.

The specific features included in the dataset are described in the accompanying file "kddcup.names". These features encompass various aspects of a network connection, including protocol type, service used, duration, number of bytes transferred, and content features (in some cases).

The KDD Cup 1999 data provides two main partitions:

\* Training data: This partition is used to train intrusion detection models. It contains a representative sample of various network connections, including both normal and attack types.

\* Test data: This partition is used to evaluate the performance of trained intrusion detection models. It includes labeled and unlabeled network connections.

The KDD Cup 1999 data is a widely used benchmark for network intrusion detection research. However, it is important to acknowledge potential limitations associated with the dataset, such as class imbalance and biases introduced during the data generation process.

1. RESULT



 Fig:02 Dashboard of system



 Fig:03 Malicious packets send

1. FUTURE SCOPE

Beyond the current implementation, the NIDS offers exciting possibilities for future development. We can delve into more advanced deep learning architectures like Convolutional Neural Networks (CNNs) or transformers to potentially achieve even more sophisticated attack pattern recognition. Additionally, exploring a hybrid approach that combines signature-based and anomaly-based detection methods holds promise for a more comprehensive security strategy. Furthermore, research into unsupervised learning techniques could equip the NIDS to identify novel or zero-day attacks not present in the training data. To enhance user trust and understanding, incorporating explainable AI (XAI) would provide valuable insights into the decision-making process of the machine learning models. Finally, real-world deployment and evaluation in a controlled network environment will be crucial to solidify the effectiveness of the NIDS against the ever-evolving landscape of cyber threats.

1. CONCLUSION

In conclusion, this paper presented a Machine Learning-powered NIDS utilizing LSTM and KNN for network traffic analysis and attack classification. The MERN stack provides a scalable architecture with a user-friendly interface and secure authentication. This NIDS offers real-time threat detection and promises enhanced security. Future avenues include exploring advanced deep learning architectures, integrating hybrid detection methods, and incorporating explainable AI for better user understanding. Real-world deployment and evaluation will further solidify the effectiveness of this NIDS against evolving cyberattacks.

1. REFERENCES

[1] Daniel Barbara, Ningning Wu and Sushil Jajodia Detecting novel network intrusion using bayes estimators. In Proceedings of First SIAM Conference on data mining Chicago, 2001.

[2] Eric Bloedorn, Alan D. Christiansen, William Hill, Clement Skorupka, Lisa M. Talbot, and Jonathan Tivel. Data mining for network intrusion detection: How to get started. Technical report, The MITRE Corporation, 2001. [3] Markus Breunig, Hans-Peter Kriegel, Raymond T. Ng, and J¨org Sander. Lof: Identifying densitybased local outliers. In Proceedings of the ACM SIGMOD Conference, Dallas, TX, 2000.

 [4] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In Proceedings of the ACM SIGMOD Conference, pages 427–438, Dallas, TX, 2000.

[5] Aleksander Lazarevic, Levent Ert¨oz, Aysel Ozgur, Vipin Kumar, and Jaideep Srivastava. A comparative study of anomaly detection schemes in network intrusion detection. In Proceedings of Third SIAM International Conference on Data Mining, San Francisco, 2003.

 [6] Elio Lozano, Edgar Acu˜na. Parallel algorithms for distance-based and density-based outliers, Data Mining, Fifth IEEE International Conference on Publication Date: 27-30 Nov. 2005.

 [7] “Deep Learning in Intrusion Detection Systems,” *ieeexplore.ieee.org*. <https://ieeexplore.ieee.org/document/8625278>

[8] M. Zaman and C. -H. Lung, "Evaluation of machine learning techniques for network intrusion detection," *NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium*, Taipei, Taiwan, 2018, pp. 1-5, doi: 10.1109/NOMS.2018.8406212.

‌[9] Almutairi YS, Alhazmi B, Munshi AA. Network Intrusion Detection Using Machine Learning Techniques. Advances in Science and Technology Research Journal. 2022;16(3):193-206. doi:10.12913/22998624/149934.

[10] Kunal and M. Dua, "Machine Learning Approach to IDS: A Comprehensive Review," *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2019, pp. 117-121, doi: 10.1109/ICECA.2019.8822120.

[11] A. Ali *et al*., "Network Intrusion Detection Leveraging Machine Learning and Feature Selection," *2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET)*, Charlotte, NC, USA, 2020, pp. 49-53, doi: 10.1109/HONET50430.2020.9322813.

[12] L. A. H. Ahmed and Y. A. M. Hamad, "Machine Learning Techniques for Network-based Intrusion Detection System: A Survey Paper," *2021 National Computing Colleges Conference (NCCC)*, Taif, Saudi Arabia, 2021, pp. 1-7, doi: 10.1109/NCCC49330.2021.9428827.

[13] R. Sommer and V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection," *2010 IEEE Symposium on Security and Privacy*, Oakland, CA, USA, 2010, pp. 305-316, doi: 10.1109/SP.2010.25.

[14] U. S. Musa, M. Chhabra, A. Ali and M. Kaur, "Intrusion Detection System using Machine Learning Techniques: A Review," *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, Trichy, India, 2020, pp. 149-155, doi: 10.1109/ICOSEC49089.2020.9215333.

[15] A. Halimaa A. and K. Sundarakantham, "Machine Learning Based Intrusion Detection System," *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 2019, pp. 916-920, doi: 10.1109/ICOEI.2019.8862784.