Big Data and Deep Learning for Scalable Network Traffic Monitoring and Analysis

V. Lakshmi Chaitanya, M. Sharmila Devi, K. Vyshnavi, U. Deepika,
S. Sana Samreen and U. Jayanthi
Department of Computer Science and Engineering, Santhiram Engineering College, Nandyal-518501, Andhra Pradesh, India

Keywords: Network Traffic, Deep Learning, Big Data, RNN, CNN, Apache Spark, Hadoop, Neural Networks.

Abstract: Network traffic analysis is important to assure the security, efficiency and proper management of digital communications. Due to the increased speed and complexity of cyberspace, traditional methods for monitoring network traffic are often difficult to enable threats. Modern network security is based on advanced data analytics techniques that handle large amounts of information to identify potential threats and anomalies. This paper focuses on implementing big data technologies and deep learning models to improve large-scale network traffic analytics. Machine learning techniques and deep learning techniques such as folding fish networks (CNNS) and repeating neural networks (RNNs) can help identify suspicious activity by determining patterns of network traffic. As the internet speed increases and more devices are created on the network, traditional methods become effective for large data flows. Using big data frameworks such as Apache Spark and Hadoop, systems can efficiently process and analyze network traffic in real time. Big data integration in deep learning improves security by quickly capturing and reducing cyber threats such as: This paper examines how these advanced technologies can enhance network security, prevent attacks, and improve overall Internet security. Additionally, the future developments and improvements in network surveillance to ensure a safer and more efficient digital environment. The ultimate goal is to create a system that protects our data, improves online security, and ensures seamless digital interaction.

1 INTRODUCTION

The rapid expansion of digital networks and the increasing reliance on internet-based transactions have led to critical concerns about network traffic monitoring and safety. Network Traffic Analysis (NTA) plays a key role in detecting malicious activity and often relies on intrusion detection systems (IDs) and various surveillance models. However, traditional methods include sensitivity to cyber threat development, scalability issues, and inefficiency in the conversion of large amounts of network traffic. For example, a signature-based ID relies on predefined attack patterns and is not effective against new or unknown threats. Furthermore, traditional approaches for machine learning need to fight and require continuous manual updates in adapting to dynamic network conditions that limit long-term effectiveness. To address these challenges, this study suggests an advanced framework for network traffic analysis using large data technologies and deep learning models to obtain more accurate and efficient intrusion recognition. Technologies such as Apache Spark and Hadoop allow large-scale data processing, while deep learning models, particularly long-term memory networks (LSTMs), improve the system's ability to recognize complex attack patterns over time. In contrast to traditional methods, deep learning models can automatically learn from a vast dataset, reducing dependency on predefined rules and improving the ability to recognize threats. Additionally, the system integrates a hybrid learning approach combining monitored techniques and unmanned techniques to improve accuracy and adaptability. By using this progress, the proposed framework improves security, real-time recognition of threats, and ensures more effective network monitoring. This study examines the implementation of these methods, the impact of cybersecurity

Chaitanya, V. L., Devi, M. S., Vyshnavi, K., Deepika, U., Samreen, S. S. and Jayanthi, U.
Big Data and Deep Learning for Scalable Network Traffic Monitoring and Analysis.
DOI: 10.5220/0013886100004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 2*, pages 538-542
ISBN: 978-989-758-777-1

impacts, and future improvements to create a safer and more resilient digital network infrastructure.

2 LITERATURE REVIEW

Hodo, E., Bellekens, X., Hamilton, A., Dubouilh, P., Iorkyase, E. & Tachtatzis, C. (2021) “Threat Analysis of IoT Networks Using Machine Learning" This study focuses on using models for machine learning, including random forests and support vector machines (SVMs) in IoT network security. We demonstrated how these models can recognize a variety of cyber threats in an IoT environment. However, it also raised concerns about scalability and the need to adapt to continuous learning models to adapt to threat development.

Lee, W. & Stolfo, S.J. In this study, we proposed a related rule for machine learning methods, especially decision-making-tree and detection of network representations. This approach showed an improved accuracy compared to traditional rule-based systems. However, effectiveness was limited by selection of characteristics requiring manual adjustment and expert intervention.

Breiman, L. (2001) “Random Forest" Breiman's work demonstrated the random forest algorithm. This was often used in network security to recognize cyber threats. Researchers used it in network traffic analysis because it can handle high-dimensional data and improve classification accuracy. Recognizing known threats, Random Forest fell smoothly, but his ability to recognize zero-day attacks was limited.

Tavallaee, M., Bagheri, E., Lu, W. & Ghorbani, A. A. (2009) "Detailed analysis of KDD Cup 99 data records." This article critically analyzed a wide range of KDDCUP99 data records, highlighting their redundancy and limitations. This has led to the development of new data records such as NSL-KDD and CICIDS2017.

Goodfellow, I., Bengio, Y. & Courville, A. (2016) “Deep Learning" This work laid the foundation for deep learning techniques applied to a variety of fields, including cybersecurity. The authors have introduced advanced neuron network architectures such as CNNS and LSTM. This has become critical when recognising networked styles. The deep learning model has proven effective against complex patterns of learning from network traffic, but requires high computational resources.

3 METHODOLOGY

The proposed methodology focuses on the integration of large data processing with deep learning techniques to improve network traffic analysis and intrusion recognition. Traditional methods make it important for scalability and adaptability to develop robust and intelligent systems that can efficiently master large amounts of network traffic. Figure 1 show the System Architecture.

Data collection and processing: A network traffic is collected from different types of thresholds, including networks (ID), and data sets), and NSL-KDD. The collected data was processed to experience the cancel of noise, handle the lost values, and the appropriate surface network compounds for the filter.

Great data processing framework: To control - scale large-network data is good; this method combines a distributing distribution as an Apache Sneak and Hadoop. This technology is controlled as preparing the scale, sending monitoring and analysis of a periodic. The use of high data framework can store, restore, and a large network list.

Removal and selection of parts: The network traffic results, such as the size of the package, protocol type, and chat, and relevant to providing great knowledge of the network. Damaged conditions such as the analysis of key areas (PCA) and autoencoder are used to optimize symbols and improves modeling. The meaning of the display indicates the identification of the time period depending on the calculations and calculations in traffic models.

Development of study patterns in -depth: The main reason of this method is the application of accessory global classic composition (CNSS) and short memory networks (LSTM): CNN is used to take language startup data from network traffic data, capturing difficult models related to normal and negative action. LSTM is used to cut the youth trust and find the long-term conditions and increases complex attack patterns. The mixed model is trained using the square basement, using methods such as preparing and restrictions to form the examples.

3.1 Architecture

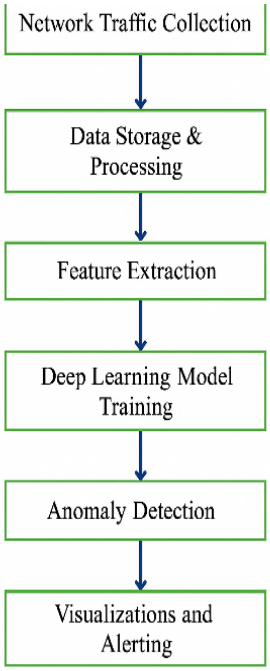

Figure 1: System Architecture.

3.2 Dataset Information

Dataset are very important for the deep learning models to analyse the network traffic. Every dataset consists of both normal and abnormal activities in the traffic. The network dataset consists of some special features like source ip address, destination ip address, source port, destination port, protocol type. Sometime based features are flow duration, packet inter arrival time and number of packets per second. Some of the attacks like deniol of service, brute force attack, web-based attack, malware and botnet attacks Some of the mathematical formulas are: Table 1 show the Network Traffic Data Set.

Packet inter arrival time: $IAT=T_{i+1}-T_i$
Where T=time
Flow duration Calculation: $D_f = T_{end} - T_{start}$
Where D_f=flow duration
T_{en} = end time
T_{start} = start time
Packet rate per second: $PRPS = N_p / D_f$
Where Df= flow duration
Packet Rate: Total Time /Interval Total Packets
Byte Rate: Total Time /Interval Total Bytes

Data Set:

Table 1: Network Traffic Data Set.

Flow Duration	Total Fwd Packets	Total Backward Packets	Fwd Packet Length Max	Bwd Packet Length Max	Flow Bytes/s	Flow Packets/s	Label
16795	10	6	1747	969	742.3619315	28.96217147	0
1860	5	1	1493	1286	1940.269246	25.19838823	0
39158	4	5	1935	1857	3889.343012	8.797657125	1
45732	2	1	1071	1463	1711.990842	22.97159368	0
12284	10	4	834	1507	2928.002992	45.95605866	0
7265	19	2	1745	872	3775.762295	41.36101835	0
17850	1	10	536	993	985.7467159	14.72231143	1
38194	5	5	1488	1250	3583.841073	25.99735031	0
22962	13	10	777	1679	913.4189947	41.75254412	0
48191	4	1	1719	790	4843.08384	8.032304965	0
45131	16	6	1415	1667	2697.773585	22.43266636	0

4 IMPLEMENTATION AND RESULTS

In this section the implementation details are mentioned to detect the network traffic. 4.1 Section contains the model selection and 4.2 section contains the results of the implements.

4.1 Model Selection

Model 1: Random Forest: An effective ensemble learning technique for intrusion detection and network traffic analysis is Random Forest. In order to ensure high accuracy and resistance to overfitting, it builds numerous decision trees and uses majority voting to classify network traffic. By processing high-

dimensional datasets and choosing key traits, it effectively manages massive amounts of data. Data collection, preprocessing, feature selection, model training, and classification of network traffic into normal or attack categories, such as DoS, DDoS, and botnet incursions, are all part of the implementation process. Prior to being implemented in real-time monitoring systems, the model is assessed using metrics for recall, accuracy, and precision. It is very effective in improving network security because of its scalability, handling of missing values, and flexibility in response to changing threats. It improves cybersecurity through real-time threat detection and mitigation when combined with big data frameworks and deep learning techniques.

Model-2: Decision Tree: Because of its ease of use and interpretability, decision trees are a popular machine learning technique for network traffic analysis and intrusion detection. It creates a tree-like structure with each node representing a decision rule by dividing data into branches according to feature requirements. Until a final prediction is produced, the classification process proceeds through a succession of logical conclusions. By examining network traffic characteristics like packet size, flow time, and protocol type, decision trees are useful for spotting attack trends. They can identify threats like DoS, DDoS, and brute force assaults and categorize traffic into benign or malevolent groups. Accuracy, precision, and recall measures are used to assess the algorithm's performance once it has been trained on labelled datasets. Nevertheless, pruning strategies or incorporating them into ensemble approaches like Random Forest might help reduce the overfitting that Decision Trees may have. They are useful for real-time network security applications because of their capacity to manage big datasets and offer transparent decision-making logic.

Model-3: Logistic Regression: Logistic regression is frequently used in conjunction with deep neural networks (DNNs) or convolutional neural networks (CNNs) in deep learning-based network security. In this case, intricate patterns are extracted from network traffic data using deep learning models, and attack probabilities are predicted using logistic regression in the last classification layer. Large datasets with characteristics like packet size, flow time, and protocol type, such as CICIDS2017 are used to train the model. Logistic Regression determines the best weights to efficiently distinguish between malicious and legitimate traffic using gradient descent and backpropagation. The capacity of logistic regression to effectively manage massive network traffic and produce probabilistic outputs that can support decision-making is one of the benefits of utilizing it in deep learning. Deep learning architectures, on the other hand, improve its capabilities by extracting hierarchical and non-linear features, as it presupposes a linear decision boundary. Threats including Denial-of-Service (DoS), botnets, and brute-force attacks can be accurately detected using deep learning models employing Logistic Regression as the last classification layer in real-time intrusion detection systems.

4.2 Results

Table 2 show the accuracy, precision, recall and F1-score values for the different algorithms are shown below. Figure 2 show the Accuracy of Deep Learning Models.

Table 2: Performance of Different Deep Learning Models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Neural Networks	90.6	84.6	95.7	89.8
Random Forest	81.1	72.4	91.3	80.8
Decision Tree	67.9	60.7	73.9	66.7
Logistic Regression	60.4	53.3	69.6	60.4

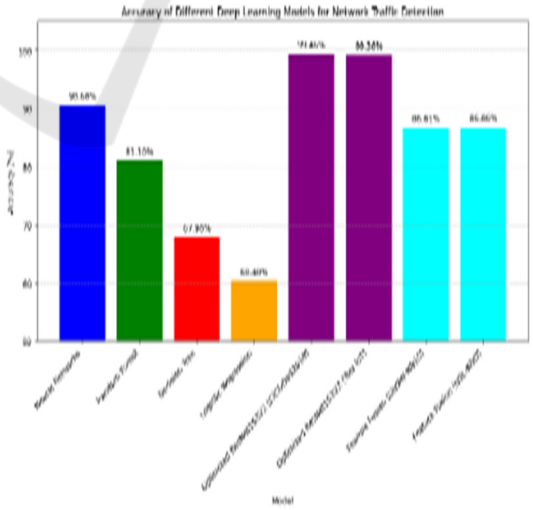

Figure 2: Accuracy of Deep Learning Models.

5 CONCLUSIONS

This paper shows the size of the data and browse a network activity to make the most of the internet wounds. Using deep study patterns such as CNSS, RNN, and the transformers, we can discard the amount of network data. Although these conditions showcase the most important promises, challenges such as major compound costs, the need for data labeled and is still in the same description. Future work needs to focus on speed, and know more, and can be learned from small data. The combination of study in the study and traditional security methods can also be better way for network protection. To briefly, the use of major and study data in the analysis of network transfer is essential to threatening and compares to the digital environment.

REFERENCES

Chaitanya, V. Lakshmi. "Machine Learning Based Predictive Model for Data Fusion Based Intruder Alert System." journal of algebraic statistics 13.2 (2022): 2477-2483

D. E. Denning, "An intrusion-detection model," IEEE Transactions on Software Engineering, vol. SE-13, no. 2, pp. 222–232, 1987.

Devi, M. Sharmila, et al. "Machine Learning Based Classification and Clustering Analysis of Efficiency of Exercise Against Covid-19 Infection." Journal of Algebraic Statistics 13.3 (2022): 112-117.

Devi, M. Sharmila, et al. "Extracting and Analyzing Features in Natural Language Processing for Deep Learning with English Language." Journal of Research Publication and Reviews 4.4 (2023): 497-502.

Jun L., Feng L., Nirwan A., "Monitoring and Analyzing Big Traffic Data of a Large-Scale Cellular Network with Hadoop", IEEE Network, Japan, vol. 28, Iss. 4, pp. 32-39, 2014.

M.Amareswara Kumar, "Baby care warning system based on IoT and GSM to prevent leaving a child in a parked car"in International Conference on Emerging Trends in Electronics and Communication Engineering - 2023, API Proceedings July-2024.

M.Amareswara Kumar,Effective Feature Engineering Technique For Heart Disease Prediction With Machine Learning" in International Journal of Engineering & Science Research, Volume 14, Issue 2, April-2024 with ISSN 2277-2685.

Mahammad, Farooq Sunar, Karthik Balasubramanian, and T. Sudhakar Babu. "A comprehensive research on video imaging techniques." All Open Access, Bronze (2019).

Mahammad, Farooq Sunar, and V. Madhu Viswanatham. "Performance analysis of data compression algorithms for heterogeneous architecture through parallel approach." The Journal of Supercomputing 76.4 (2020): 2275-2288.

Mahammad, Farooq Sunar, et al. "Key distribution scheme for preventing key reinstallation attack in wireless networks." AIP Conference Proceedings. Vol. 3028. No. 1. AIP Publishing, 2024.

Mandalapu, Sharmila Devi, et al. "Rainfall prediction using machine learning." AIP Conference Proceedings. Vol. 3028. No. 1. AIP Publishing, 2024.

Mohd R. G., Durgaprasad G., "Hadoop, MapReduce and HDFS: A Developers Perspective", in International Conference on Intelligent Computing, Communication & Convergence ((ICCC-2014), India, pp.45-50, 2015.

Paradesi SubbaRao," Morphed Image Detection using Structural Similarity Index Measure"M6 Volume 48 Issue 4(December 2024, https://powertechjournal.com

Paradesi Subba Rao,"Detecting malicious Twitter bots using machine learning" AIP Conf. Proc. 3028, 020073 (2024),https://doi.org/10.1063/5.0212693

Parumanchala Bhaskar, et al. "Machine Learning Based Predictive Model for Closed Loop Air Filtering System." Journal of Algebraic Statistics 13.3 (2022): 416-423.

Parumanchala Bhaskar, et al. "Incorporating Deep Learning Techniques to Estimate the Damage of Cars During the Accidents" AIP Conference Proceedings. Vol. 3028. No. 1. AIP Publishing, 2024.

Parumanchala Bhaskar, et al "Cloud Computing Network in Remote Sensing-Based Climate Detection Using Machine Learning Algorithms" remote sensing in earth systems sciences(springer).

R. Fontugne, J. Mazel, K. Fukuda, "Hashdoop: A MapReduce Framework for Network Anomaly Detection", in 2014 IEEE INFOCOM Workshops: 2014 IEEE INFOCOM Workshop on Security and Privacy in Big Data, Japan, pp. 494-499, 2014.

Suman, Jami Venkata, et al. "Leveraging natural language processing in conversational AI agents to improve healthcare security." Conversational Artificial Intelligence (2024): 699-711.

Sunar, Mahammad Farooq, and V. Madhu Viswanatham. "A fast approach to encrypt and decrypt video streams for secure channel transmission." World Review of Science, Technology and Sustainable Development 14.1 (2018): 11-28.

W. Lee and S. J. Stolfo, "A data mining framework for building intrusion detection models," in Proceedings of the 2000 IEEE Symposium on Security and Privacy (S&P), Oakland, CA, USA, 2000, pp. 120–132.