www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 883-886

e-ISSN : 2583-1062
Impact Factor : 7.001

ENHANCING CYBERSECURITY THREAT DETECTION USING DEEP LEARNING ALGORITHMS ON BIG DATA

Meenu Sharma[1], Jyoti Rana[2]

[1]Assistant Professor Department of Artificial Intelligence & Data Science, Dr. Akhilesh Das Gupta Institute of Professional Studies, Delhi, India.

[2]Scholar of B. Tech 3rd Year Department of Artificial Intelligence & Data Science, Dr. Akhilesh Das Gupta Institute of Professional Studies, Delhi, India.

meenu.kodnya@gmail.com

jyotiserana@gmail.com

DOI : https://www.doi.org/10.56726/IRJMETS36820

ABSTRACT

This paper explores the potential of deep learning algorithms in analyzing vast amounts of cybersecurity data to detect and mitigate threats. The proposed approach employs convolutional neural networks (CNNs) and long short-term memory (LSTM) networks to classify and predict malicious activity with high accuracy. This paper presents a novel architecture that combines CNN and LSTM layers, offering improved feature extraction and temporal analysis.

Keywords- Deep Learning, Cybersecurity, Big Data, Threat Detection, Convolutional Neural Networks, Long Short-Term Memory

1. INTRODUCTION

The rise of digital interactions has also introduced significant security challenges. Cybersecurity [1] threats are evolving, and traditional approaches often fall short in detecting complex, multi-dimensional attacks. Given the volume of data generated in real-time from network activity and connected devices, deep learning approaches are promising due to their ability to handle complex, high-dimensional data and identify patterns associated with cybersecurity threats.

Traditional methods often use rule-based systems, which are efficient but struggle with new, previously unseen types of attacks. While Machine Learning (ML) [2] techniques are more adaptable, they often require significant feature engineering, which deep learning [3] can mitigate by autonomously learning feature representations from raw data.

This paper aims to design and evaluate deep learning architectures to improve threat detection accuracy in cybersecurity. Objectives include comparing CNN [4] –LSTM [5] hybrid models against standard ML techniques and evaluating performance on large datasets.

2. LITERATURE REVIEW

The integration of deep learning into cybersecurity represents a significant advancement in how we approach threat detection and prevention. This literature review explores various studies and methodologies that showcase the application of deep learning techniques in cybersecurity, highlighting their effectiveness and identifying areas for future research.

A study by [15] proposed an innovative combined deep learning approach to tackle the challenges of detecting pirated software and malware-infected files within IoT [6] networks. This approach utilizes TensorFlow's deep neural network to identify source code plagiarism, employing tokenization and weighting feature methods to filter noisy data and enhance the importance of each token in terms of source code. This method improves the detection of source code plagiarism by focusing on key features and patterns. Additionally, the study integrates a deep convolutional neural network (CNN) to detect malicious infections by analyzing color image visualizations of the IoT network. This dual approach highlights the versatility of deep learning techniques in addressing different types of cybersecurity threats.

In [16], the authors presented a comprehensive survey of deep learning approaches applied to cybersecurity intrusion detection. The paper reviews various deep learning models, including recurrent neural networks (RNNs) [7], deep neural networks (DNNs) [8], restricted Boltzmann machines (RBMs) [9], deep belief networks (DBNs) [10], convolutional neural networks (CNNs), deep Boltzmann machines (DBMs) [11], and deep autoencoders. It also emphasizes the critical role of datasets in intrusion detection and provides a classification of 35 well-known cyber datasets into seven categories: network traffic based, electrical network based, internet traffic based, virtual private network based, android apps based, IoT traffic based, and internet connected devices based datasets. This survey offers valuable insights into the diverse applications of deep learning models and their effectiveness across different types of cybersecurity datasets.

The study by [17] explores the application of machine learning techniques in monitoring and analyzing cybersecurity threats in cloud environments, with a focus on enterprise applications in telecommunications and IoT. The paper proposes combining Support Vector Machines (SVMs) [12], neural networks, and deep neural networks (DNNs) to enhance threat detection capabilities. It introduces an approach for aggregating classifier results based on performance weights, which has shown promising results comparable to existing algorithms. This approach highlights the potential of combining multiple machine learning techniques to improve security applications in complex and dynamic environments.

In [18], the authors provide an analysis of machine learning techniques used for detecting intrusions, malware, and spam. The study aims to assess the current maturity of these solutions and identify their limitations. Based on an extensive review of the literature and experiments conducted on real enterprise systems and network traffic, the paper outlines the strengths and weaknesses of various machine learning approaches. This analysis is crucial for understanding the practical challenges and limitations of adopting machine learning-based cybersecurity solutions in real-world scenarios.

The research presented in [19] explores how machine learning can enhance cyber threat detection through behavioral modeling of network traffic patterns. By using anomaly detection based on machine learning, this study provides adaptive protection by learning normal behavior and identifying deviations that may indicate malicious activity. The paper covers a range of machine learning techniques, including supervised, unsupervised, and hybrid algorithms such as neural networks, support vector machines (SVMs), random forests [13], self-organizing maps, k-means clustering, and isolation forests. It highlights the ability of these techniques to model complex patterns in network traffic data that are not discernible by traditional rule-based methods, offering valuable guidance for maximizing detection capabilities.

3. METHODOLOGY

A. Dataset

For this study, we use the “cyberfeddefender_dataset” dataset, which 1,430 instances, with 23 features including information on packet size, duration, bytes sent/received, flow statistics, and attack labels. It covers common cyberattacks along with normal network traffic. The dataset was split into training (80%) and testing (20%) sets, and features were normalized.

- **Data Loading**: This involves importing data, here we use “cyberfeddefender_dataset” CSV file.
- **Data Cleaning**: This step ensures that the data is free from errors and inconsistencies. It includes:
  - Removing duplicates
  - Handling missing values
- **Preprocessing**: This step prepares the data for analysis and modeling. It includes:
  - **Encoding Categorical Variables**: Converting categorical data into numerical format using techniques label encoding.
  - **Standardization**: Transforming data to have a mean of 0 and a standard deviation of 1.

B. Model Architecture

The proposed model is a CNN-LSTM hybrid network. The CNN layers handle spatial feature extraction, while the LSTM layers process sequential patterns, enhancing the detection of time-dependent anomalies.

- **Conv1D Layers:**
  - Extract spatial features from network traffic.
  - Apply convolution operations along one dimension, useful for processing time-series data or sequences.
  - Identify patterns and anomalies by analyzing the sequence of data packets.
  - Enhance the model's ability to detect subtle changes in network traffic.
- **MaxPooling Layers:**
  - Reduce feature dimensionality to prevent overfitting.
  - Down-sample the data by taking the maximum value from a set of values within a small window.
  - Make the model more efficient by reducing the number of parameters and computations.
  - Help the model generalize better by focusing on the most important features.
- **LSTM Layers:**
  - Process the sequential nature of network events.

- Capture long-term dependencies and patterns in the data.
- Handle sequences where the order of events is important.
- Remember information over long periods, reducing the vanishing gradient problem.

- **Dense Layers:**
- Output the final classification decision.
- Use a sigmoid activation function for binary classification (normal vs. anomalous).
- Convert the processed features into a probability value between 0 and 1.
- Provide the final decision on whether the network traffic is normal or anomalous.

C. Training & Evaluation

The model was trained using loss function "binary cross-entropy" and the optimizer "Adam" with an initial learning rate of 0.0005. The dropout layers were applied to reduce overfitting. Evaluation of the model is done by checking the accuracy.

4. RESULT

The model achieved an average accuracy of 93%, surpassing baseline machine learning models such as logistic regression and decision trees. The CNN-LSTM model showed superior performance in identifying both high-frequency, less complex attacks and low-frequency, complex attacks. However, some challenges remain in optimizing real-time processing speed and reducing model complexity.

5. CONCLUSION

This research demonstrates the value of deep learning, particularly CNN-LSTM hybrid models, for improving cybersecurity threat detection. By leveraging big data and deep learning, this approach could significantly enhance the accuracy of real-time threat detection systems.

ACKNOWLEDGEMENT

I, Jyoti Rana would like to express my sincere gratitude to all those who have contributed to the successful completion of this project. Special thanks to Ms. Meenu Sharma for their invaluable guidance and support throughout the process. I also wish to acknowledge the assistance provided by Dr. Akhilesh Das Gupta Institute of Professional Studies and its staff, whose resources and expertise were instrumental in achieving the project's objectives.

Additionally, I am grateful to my family and friends for their unwavering encouragement and understanding during this journey. Their support has been a constant source of motivation.

6. REFERENCES

[1] What Is Cybersecurity? | IBM (https://www.ibm.com/topics/cybersecurity)
[2] What Is Machine Learning (ML)? | IBM (https://www.ibm.com/topics/machine-learning)
[3] What Is Deep Learning? | IBM (https://www.ibm.com/topics/deep-learning)
[4] What are Convolutional Neural Networks? | IBM (https://www.ibm.com/topics/convolutional-neural-networks)
[5] What is LSTM - Long Short Term Memory? - GeeksforGeeks (https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/)
[6] Introduction to Internet of Things (IoT) - Set 1 - GeeksforGeeks (https://www.geeksforgeeks.org/introduction-to-internet-of-things-iot-set-1/)
[7] What is a Recurrent Neural Network (RNN)? | IBM (https://www.ibm.com/topics/recurrent-neural-networks)
[8] What's a Deep Neural Network? Deep Nets Explained – BMC Software | Blogs (https://www.bmc.com/blogs/deep-neural-network/)
[9] Restricted Boltzmann Machine - GeeksforGeeks (https://www.geeksforgeeks.org/restricted-boltzmann-machine/)
[10] Deep Belief Network (DBN) in Deep Learning - GeeksforGeeks (https://www.geeksforgeeks.org/deep-belief-network-dbn-in-deep-learning/)
[11] Deep Boltzmann Machines (DBMs) in Deep Learning - GeeksforGeeks (https://www.geeksforgeeks.org/deep-boltzmann-machines-dbms-in-deep-learning/)
[12] How SVM Works - IBM Documentation (https://www.ibm.com/docs/en/spss-modeler/saas?topic=models-how-svm-works)
[13] What Is Random Forest? | IBM (https://www.ibm.com/topics/random-forest)

[14] What is Max pooling in CNN? is it useful to use? | by Rahul Kadam | CodeX | Medium (https://medium.com/codex/what-is-max-polling-in-cnn-is-it-useful-to-use-6f2d6ff44c6)

[15] Farhan Ullah, Hamad Naeem, Sohail Jabbar ,Shehzad Khalid, Muhammad Ahsan Latif, Fadi Al-Turjman, And Leonardo Mostarda, "Cyber Security Threats Detection in Internet of Things Using Deep Learning Approach", Digital Object Identifier 10.1109/ACCESS.2019.2937347 (2019)

[16] Mohamed Amine Ferrag, Leandros Maglaras, Sotiris Moschoyiannis , Helge Janicke, "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study", in Journal of Information Security and Applications 50 (2020) 102419

[17] S. A. Sokolov, T. B. Iliev and I. S. Stoyanov, "Analysis of Cybersecurity Threats in Cloud Applications Using Deep Learning Techniques", MIPRO 2019/CTI

[18] Giovanni Apruzzese, Michele Colajanni, Luca Ferretti, Alessandro Guido, Mirco Marchetti, "On the Effectiveness of Machine and Deep Learning for Cyber Security", in Cyber Law and Espionage Law as Communicating Vessels (indiana.edu), Page No. 371

[19] Fatima Bouchama, Mostafa Kamal Internationa, "Enhancing Cyber Threat Detection through Machine Learning-Based Behavioral Modeling of Network Traffic Patterns", International Journal of Business Intelligence and Big Data Analytics