AI-Powered Intrusion Detection System for Network Security Using Big Data

Jagriti Kumari
Department of Computer Science
Chandigarh University
Mohali, Punjab, India
Kri.jagriti.79@gmail.com

Himanshu
Department of Computer Science
Chandigarh University
Mohali, Punjab, India
dhimanshanu07@gmail.com

Ashutosh Sharma
Department of Computer Science
Chandigarh University
Mohali, Punjab, India
ash067142@gmail.com

Anjali Kumari
Department of Computer Science
Chandigarh University
Mohali, Punjab, India
dvuanjali11@gmail.com

Khushi Kumari
Department of Computer Science
Chandigarh University
Mohali, Punjab, India
khushipbps123@gmail.com

Shaffy
Department of Computer Science
Chandigarh University
Mohali, Punjab, India
shaffy18@gmail.com

Abstract— Organisations are increasingly using artificial intelligence (AI) to strengthen their cybersecurity defences, especially against phishing attempts and intrusion detection, as cyber threats continue to change. A significant development in this field is represented by AI-powered supervised classifiers, which utilize vast amounts of data to enhance detection accuracy and response times. This study investigates the use of AI-driven supervised classifiers in big data settings, emphasising how well they can detect intrusions in real time and identify and mitigate phishing risks.

This study investigates AI-based threat detection methods, assesses their efficacy in cybersecurity, and discusses the moral ramifications of AI-powered security solutions. Future cybersecurity frameworks can improve resilience against cyber threats and guarantee strong digital protection in a period of growing cyber hazards by combining AI with blockchain technology, quantum computing, and federated learning.

Keywords: Decentralised Structure, Quantum Honeypot, Command-Based Content Problem, SSH/Telnet, Natural Language Processing (NLP), Supervised Learning, Signature, Autoencoders, Anomaly Detection.

I. INTRODUCTION

By automating threat detection, lowering the need for human intervention, and improving reaction capabilities, AI-driven cybersecurity systems provide notable benefits over traditional security measures. [1]Security frameworks can detect departures from typical behaviour by using machine learning models, such as supervised learning classifiers and unsupervised anomaly detection algorithms, which indicate possible threats before they become more serious [1] [2] [3]. By evaluating enormous volumes of unstructured data, spotting attack trends, and enhancing malware categorisation, deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), further improve security [4] [3].

Integrating blockchain technology with AI can improve security and transparency while fortifying AI-driven cybersecurity frameworks [5]. Attackers find it challenging to alter security logs or introduce malicious code into AI-based systems due to the blockchain's decentralized structure, which ensures data integrity [5]. Furthermore, developments in quantum computing 12 present cybersecurity with both potential and risks, requiring the use

of quantum-resistant encryption methods to protect AI-driven security systems [6].

This study examines AI-based cybersecurity methods, assesses how well they detect threats, and discusses potential uses in the future to improve digital security. Organisations may create stronger cybersecurity plans and guarantee defence against ever-more-sophisticated cyberthreats in the digital era by tackling present issues and incorporating AI with cutting-edge technologies.

Machine learning, deep learning, and natural language processing are some of the cutting-edge technologies included in the study that give systems the ability to anticipate, identify, and stop threats in real time [1] [3] [7] [8]. Because artificial intelligence can process vast amounts of data and identify patterns, it is a potent weapon in cybersecurity. By enhancing the ability to recognise threats, technology is used to improve cybersecurity.

II. LITERATURE REVIEW

First, signature-based methods were used to identify known threats. New security vulnerabilities introduced by technological advancements make it challenging for outdated intrusion detection methods to function effectively. Consequently, AI-powered solutions were implemented to tackle the security concerns [1] [3]. One use of machine learning techniques is the monitoring or identification of cloud security threats [9].

Machine learning (ML) techniques employed in intrusion detection systems for monitoring and detection are practical solutions to the security problems with cloud platforms. The researchers determined that unsupervised learning was the best machine learning method for network anomaly detection since it could discover threats that were not yet known [2] [10]. The authors claim that autoencoders are a specific type of neural network that depends on unsupervised machine learning to operate. The two primary parts of autoencoders are the encoder and the decoder [10]. The encoder part uses data as input to do compression, while the decoder part generates an output during the reconstruction phase.

Based on anomaly-based methods, the authors' investigation successfully identified previously unknown dangers when appropriately connected to cloud platforms and other systems. To demonstrate the efficacy of autoencoders in identifying intrusions, they carried out a study. High attack levels can be identified with the aid of an intrusion detection system based on auto encoders [10]; the authors found. Because of their innovative research, autoencoders are now employed as one of the network intrusion detection methods. Autoencoders are useful for monitoring and detecting network threats, but their positive rates are modest.

Attacks in the cybersecurity domain have changed significantly in terms of both quantity and type. Ransomware and phishing attacks are becoming more frequent in organizations [7] [8]. Internet Security Threat Report, which calls for the creation of more advanced and flexible cybersecurity defences. The threats and the methods used to counter them are always changing. The application of artificial intelligence to identify and address dangers is growing [1] [3] [8]. Machine learning and deep learning are becoming popular in the cybersecurity field and are referred to as game-changers.

Thanks to these technologies, systems can now gather enormous volumes of data in real-time, making it easier to identify signs of a security breach.

III. METHODOLOGY

The project's goal was to create an Intrusion Detection System (IDS) with AI capabilities that would aid in defending computer networks from online threats [4] [3] [9]. First, to train our system, we required a large amount of real-world-like data. We did this by using a publicly accessible dataset that contains comprehensive data about various network activities [4] [8], some of which are examples of hacking or unauthorized access, and others of which are entirely legitimate. To help the system learn to distinguish between good and poor behavior, think of it as providing examples of both. Cleaning the data was a crucial next step after

obtaining the dataset [4] [9]. Raw data is typically messy; it may contain fields that aren't relevant, missing values, or inconsistent formats.

Therefore, we took care to exclude or correct anything that would cause the model to become confused. Additionally, we needed to transform the data into a computer-readable format. We converted any text into numerical values and normalized all of the numbers [4] [8] because machine learning models perform best with numerical values. This ensures that all of the values are on the same scale and that large numbers do not overwhelm smaller ones. Following data preparation and cleaning, we carried out a process known as feature selection. This indicates that we selected only the most valuable columns [4] from the data that genuinely aid in the model's decision-making. For instance, knowing the weather isn't useful if you're attempting to identify a break-in. Training the machine learning model was the next big step. To determine which model produced the greatest results, we experimented with a variety of models, including Random Forest, Support Vector Machines, and Neural Networks [4] [8] [3]. In order to teach the model to recognize patterns, training essentially entails providing it with numerous instances of both benign and malevolent behavior.

For training, we used fresh data that the model had never seen before to test it. This enabled us to test the model's ability to identify dangers in practical settings. With an accuracy of 98.54%, one of the models did incredibly well and was accurate nearly all of the time [8].

Next, we developed a Python function that enables us to test this model with any fresh data [4] [9]. This function receives a network activity sample, applies the previously performed data processing steps (such as formatting and scaling), and then sends the data to the trained model to obtain a prediction

The model indicates "Intrusion detected" if it detects anything suspicious and "No threat detected" if it detects nothing dangerous [4] [9] [8]. The model also attempts to identify the kind of threat it identifies, such as data theft, denial-of-service attacks, or other threats. It's interesting to note that in one of our test instances, the model identified an incursion but classified it as "normal"; this indicates that it's functioning, but more improvement in precision is still possible.

We used Python, a robust and user-friendly programming language, to create the entire system. We made use of well-known data science packages such as NumPy [4] for working with arrays and numerical data, Scikit-learn for creating and training machine learning models, and Pandas for processing data. In the future, we want to use big data tools like Hadoop or Apache Spark to enhance this system to manage enormous volumes of real-time data [7] [9] [8]. This would make our technology extremely helpful for businesses, organizations, and government networks that require robust cybersecurity, since it would enable it to continuously monitor big networks and react to attacks immediately.

IV. RESULT

The KDD dataset was used to assess the intrusion detection model's performance [4] [8]. The dataset was preprocessed by dividing it into training and testing sets, standardizing numerical properties, and encoding categorical features. The processed data was used to train a Random Forest Classifier with 50 estimators [8].

Model Performance:

The model successfully distinguished between normal and intrusion activities on the test set, achieving an accuracy of {accuracy * 100:.2f}%.

The model's capacity to generalize to new data was enhanced by the application of feature normalization and category encoding.

The accuracy_score function from the sklearn.metrics module, which assesses the percentage of properly predicted instances in the test set, was used to calculate the accuracy

score. With an accuracy of 98.54%, the model demonstrated its ability [8] to accurately categorize network traffic as either intrusion or normal.

Intrusion Detection:

The trained model was used to assess a test instance. The instance could be accurately classified by the system as either regular behavior or an intrusion.

A sample input from the test dataset was also used to assess the trained model. After running the sample through the trained model, the intrusion detection function identified whether it was typical network traffic or an intrusion. The program proved its capacity to categorize different types of network breaches by accurately identifying possible risks.

Additionally, inverse label transformation was used to get the projected attack category, which made it possible to identify particular incursion kinds [4].

Observation:

i) The model offers an interpretable method of identifying intrusion types, making it a useful tool for network security monitoring [4] [8].

ii)The Random Forest Classifier performed well in detecting intrusions, probably because of its ensemble nature, which reduces overfitting and improves classification robustness [8].

iii)Normalization of features helped improve model convergence and performance [4] [8].

iv)Label encoding was required to convert categorical network attributes into numerical values appropriate for machine learning models [4].

To increase detection accuracy, more research might test various machine learning models, adjust hyperparameters, and assess feature importance.

V. CONCLUSION

To improve network security, this study used a Random Forest Classifier to develop an AI-powered intrusion detection system [4] [8]. The findings show that network intrusions may be efficiently detected and categorized using machine learning approaches [1] [3], especially ensemble learning. The model's performance was greatly enhanced by the preprocessing stages, which included feature encoding and normalization.

All things considered, this study demonstrates how machine learning may be used in cybersecurity [1] [3] [8] to provide a scalable and automated method of threat detection in network environments.

VI. FUTURE WORK

To improve the capabilities and adaptability of the suggested intrusion detection system, this study might be conducted in a number of ways in subsequent rounds. To manage massive, real-time network traffic, one significant improvement is combining the system with big data platforms like Hadoop or Apache Spark. This would make it possible for the IDS to more efficiently monitor and react to threats in high-volume, dynamic situations. Furthermore, adding deep learning models like CNN, LSTM, or transformer-based architectures may increase the accuracy of detection, especially for intricate, dynamic, or until undiscovered attack patterns. The use of online learning strategies, which enable the system to continually adapt to new threats without necessitating total retraining, is another interesting topic.

References

[1] M. S. Aslam, A. Alshdadi, M. A. Jan and M. Usman, “Cyber Threat Detection Using Machine Learning Techniques: A Review,” *IEEE Access,* p. 27536–27557, 2022.

[2] A. Zolanvari, M. Teixeira, L. Gupta and M. M. A. Kazim, “Machine Learning-Based Network Vulnerability Analysis of Industrial Internet of Things,” *IEEE*

Internet of Things Journal, vol. 9, no. 12, p. 9047–9058, 2022.

[3] R. Vinayakumar, K. P. Soman and P. Poornachandran, "Evaluating Deep Learning Approaches to Detect Intrusions in SDN Networks," *IEEE Communications Surveys & Tutorials,* vol. 24, no. 1, p. 658–689, 2022.

[4] Nathan Shone, Tran Nguyen Ngoc, Vu Dinh Phai and Qi Shi, "A Deep Learning Approach for Network Intrusion Detection System," *IEEE Transactions on Emerging Topics in Computational Intelligence,* vol. 6, no. 2, p. 219–229, 2022.

[5] Jayesh Rane, Saurabh P. Choudhary and Nitin Liladhar Rane, "Blockchain and Artificial Intelligence Integration for Network Security: Recent Advances and Future Directions," *IEEE Access,* vol. 10, pp. 111348-111365, 2022.

[6] A. K. Sangaiah and N. Bharti, "Quantum Computing Impact on Cybersecurity and its Defense Using AI Models," *IEEE Transactions on Industrial Informatics,* vol. 19, no. 3, pp. 3456-3465, 2023.

[7] M. A. Ferrag, L. Maglaras and A. Argyriou, "Deep Learning for Cybersecurity in IoT Networks: Threats and Solutions," *IEEE Internet of Things Magazine,* vol. 5, no. 1, pp. 50-55, 2022.

[8] S. M. Murad and Md. Ziaur Rahman, "AI-Driven Threat Detection: A Comparative Study of Deep Learning Methods," *IEEE Transactions on Network and Service Management,* vol. 20, no. 1, pp. 546-560, 2023.

[9] M. Hossain, M. F. Zolkipli and A. Y. Bin Zomaya, "Cloud-based Intrusion Detection System Using Machine Learning Algorithms," *IEEE Transactions on Cloud Computing,* 2023.

[10] J. D. Martínez-Muñoz, F. Díaz-Cano and J. A. Orte, "Autoencoders in Anomaly Detection: Application in Cybersecurity," *IEEE Access,* vol. 11, pp. 15632-15645, 2023.