IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Predictive Modelling For Network Threat Detection Using Artificial Intelligence Techniques

[1]Mr.Selventhiran T, [2] Mr.Dhanasekaran M, [3] Mr.Pranav Muthu Kumaran M,
[4] Mr.Muthu Thiruvenkadam U,[5] Mr.P.Sachudhanandam
[*1,2,3,4]Student BTech in Artificial Intelligence and Data Science, Anand Institute of Higher Technology, Chennai, Tamil Nadu, India
[5]Asst.Professor, Artificial Intelligence and Data Science, Anand Institute of Higher Technology, Chennai, Tamil Nadu, India

Abstract: The integration of artificial intelligence (AI) into cybersecurity has significantly converted how associations approach trouble discovery and forestallment. Traditional styles frequently fall suddenly when dealing with sophisticated, fleetly evolving cyber pitfalls. This paper introduces a new prophetic modelling approach that leverages AI ways to descry and alleviate implicit pitfalls in real time, offering a more dynamic and intelligent result to ultramodern network security challenges.
The proposed system utilizes machine literacy algorithms to dissect vast volumes of network business data, relating patterns that signify vicious conditioning similar as intrusions, malware propagation, and anomalies. By employing supervised literacy ways and continuously streamlining its models with new data, the system can directly read vulnerabilities and descry pitfalls before they escalate into full- scale attacks.

This visionary approach enables briskly responses and further informed decision- timber. Through the emulsion of prophetic analytics and artificial intelligence, this exploration aims to establish a scalable, adaptive, and robust cybersecurity frame. The system not only enhances real- time trouble discovery but also contributes to long- term network adaptability by minimizing homemade intervention and optimizing resource application. Eventually, this work paves the way for the wide relinquishment of intelligent trouble discovery systems in both public and private sectors, icing stronger digital security in an decreasingly connected world.

Keywords- Predictive modelling, Network security, Artificial intelligence, Threat detection, Cybersecurity, Machine learning, Anomaly detection, Predictive analytics, Network traffic analysis, Cyber threats.

I. INTRODUCTION

Data science is an interdisciplinary field that combines scientific methods, processes, algorithms, and systems to extract knowledge and insights from both structured and unstructured data. Initially proposed in 1974 by Peter Naur as an alternative to computer science, the term "data science" evolved over time, gaining formal recognition in 2008 through the efforts of D.J. Patil and Jeff Hammerbacher. The field has rapidly grown in importance, driven by its ability to generate actionable insights for decision-making in various industries. Data scientists, who combine domain expertise, programming skills, and mathematical and statistical knowledge, play a pivotal role in analyzing and managing large datasets
.Key skills for data scientists include proficiency in programming languages such as Python, SQL, and R, along with expertise in machine learning, data visualization, and big data platforms. The application of data science extends to diverse areas such as business strategy, healthcare, finance, and urban planning.
Artificial intelligence (AI), a field closely related to data science, focuses on simulating human intelligence in machines. AI systems are designed to perceive their environment, learn from it, and make decisions to achieve specific goals. AI applications, such as natural language processing (NLP), machine learning, and computer vision, have seen significant advancements, transforming industries ranging from e-commerce to

autonomous driving. AI research has evolved over decades, incorporating various approaches, from expert systems to statistical machine learning, with the goal of achieving general intelligence.

Natural language processing (NLP), a subfield of AI, enables machines to understand and interpret human language. NLP applications, such as text mining and machine translation, leverage advanced statistical methods and deep learning techniques to improve accuracy in language processing tasks. The field has made significant strides with transformer-based architectures that enable the generation of coherent and contextually accurate text.

This paper explores the intersection of data science and AI, highlighting the tools, techniques, and applications that shape these fields. We aim to provide a comprehensive overview of the current state of research and its implications for solving complex, real-world problems.

II. LITERATURE SURVEY

The paper "Malware Classification and Composition Analysis: A Survey of Recent Development" by Abusitta, Li, and Fung (2021) provides a detailed survey of existing malware classification techniques, composition analysis methods, feature extraction strategies, and malware evasion tactics. It focuses on categorizing previous research, identifying patterns, and outlining challenges in understanding malware functionalities and attacker intentions. While the survey offers valuable insights into the state of malware analysis up to 2021, it primarily remains a review without proposing new detection frameworks or addressing the latest adversarial threats and propagation models.

In our project, we build upon this foundation by advancing beyond a survey approach to develop and implement improved machine learning models, such as enhanced Convolutional Neural Networks (CNNs) and hybrid CNN-SVM frameworks, for real-time malware classification. Furthermore, we incorporate adversarial malware generation techniques, inspired by generative adversarial networks (GANs), to evaluate model robustness against sophisticated cyber attacks. Additionally, we explore malware propagation in wireless networks using modern hypergraph-based modeling, providing a dynamic analysis that was not covered in the original survey. Thus, our work not only synthesizes prior findings but also introduces practical advancements in classification accuracy, resilience against adversarial evasion, and modeling of malware spread in complex environments.

The paper "Mal Fox: Camouflaged Adversarial Malware Example Generation Based on Conv-GANs Against Black-Box Detectors" by Zhong and Cheng (2023) introduces a Conv-GAN-based framework, Mal Fox, designed to generate adversarial malware examples capable of evading third-party black-box malware detectors. Mal Fox innovatively employs three perturbation methods — Obfusmal, Steal mal, and Hollow mal — to create camouflaged malware variants, achieving high evasion rates while maintaining functional integrity. Their evaluation demonstrates significant success, with Mal Fox reducing detection rates by 45.1% and improving evasive performance by up to 56.0% over 12 baseline models. While Mal Fox represents a major step in adversarial malware generation, it primarily focuses on attacking existing black-box models without addressing broader defense mechanisms, adaptive countermeasures, or real-time adversarial robustness.

In our project, we advance beyond Mal Fox by not only generating adversarial malware examples but also integrating adversarial training into malware classification systems to improve model resilience. Additionally, we enhance the GAN-based generation process with dynamic perturbation strategies based on real-time detector feedback, aiming to simulate evolving cyber attack behaviors more realistically. Our work also extends the evaluation metrics beyond simple evasion rates to include system-level impacts, adaptive detection rates over time, and cross-model generalization, providing a more comprehensive framework for securing malware detection systems against adversarial threats.

The paper "Modelling and Analysing Malware Propagation over Wireless Networks Based on Hypergraphs" by Chen et al. (2023) proposes a hypergraph-based model to study malware spread across large-scale wireless networks. Their model captures the unique characteristics of wireless communication, particularly limited-range and Internet-independent transmissions. Using a heterogeneous mean-field approach, they derive the malware outbreak threshold and demonstrate through simulations that factors such as the number of connected devices and heterogeneous device distributions critically influence malware pandemics. Their findings also show that isolating Internet connections alone is insufficient to prevent malware outbreaks in wireless environments, highlighting the distinct vulnerability of wireless networks. While this work provides valuable theoretical insights into the dynamics of wireless malware spread, it primarily focuses on static network structures and generalized malware behaviours.

In our project, we advance beyond this model by introducing dynamic hypergraph structures that evolve based on device mobility, connection changes, and real-time network behaviour. We also incorporate different classes of malware (e.g., worms, ransomware, spyware) with varying propagation strategies to better simulate real-world cyber attacks. Furthermore, we integrate machine learning-based detection models into the network simulation to study proactive defence mechanisms against malware outbreaks, providing a more realistic and comprehensive framework for analysing and mitigating malware propagation in next-generation wireless systems.

III. SYSTEM ARCHITECTURE

The system architecture for network threat prediction is structured using Python and Django, adhering to the Model-View-Template (MVT) framework to ensure modularity and scalability. It integrates multiple machine learning algorithms, including Random Forest, Bayesian Network, and AdaBoost, which are trained on preprocessed network threat data and serialized for efficient deployment. The backend is responsible for managing user authentication, data processing, and generating threat predictions, while the frontend, developed using HTML, CSS, and JavaScript, provides an intuitive interface for data input and visualization of results. The system is designed with scalability in mind, facilitating the continuous integration of updates to both the machine learning models and underlying infrastructure. This architecture supports the effective detection of network threats and ensures adaptability to evolving cybersecurity challenges.

3.1 Overall Architecture

The proposed system for network threat prediction is built using Python and Django, leveraging the Model-View-Template (MVT) architecture to facilitate efficient and scalable data processing. The backend integrates multiple machine learning models, including Random Forest, Bayesian Network, and AdaBoost, which are trained on a Kaggle-hosted network threat dataset. These models are subjected to rigorous evaluation using performance metrics such as accuracy, precision, and recall, and the best-performing model is subsequently serialized into a .pkl file for deployment. SQLite serves as the database solution for storing user information, input data, and the corresponding prediction results. On the frontend, a dynamic user interface is developed using HTML, CSS, and JavaScript, providing interactive pages for authentication, data input, and visualization of threat predictions. The system's security is enhanced by Django's built-in mechanisms, ensuring protection against common web vulnerabilities. For deployment, the system is designed to run on Django's production-ready server with the potential to scale to cloud platforms such as AWS or Azure, ensuring reliability and flexibility.

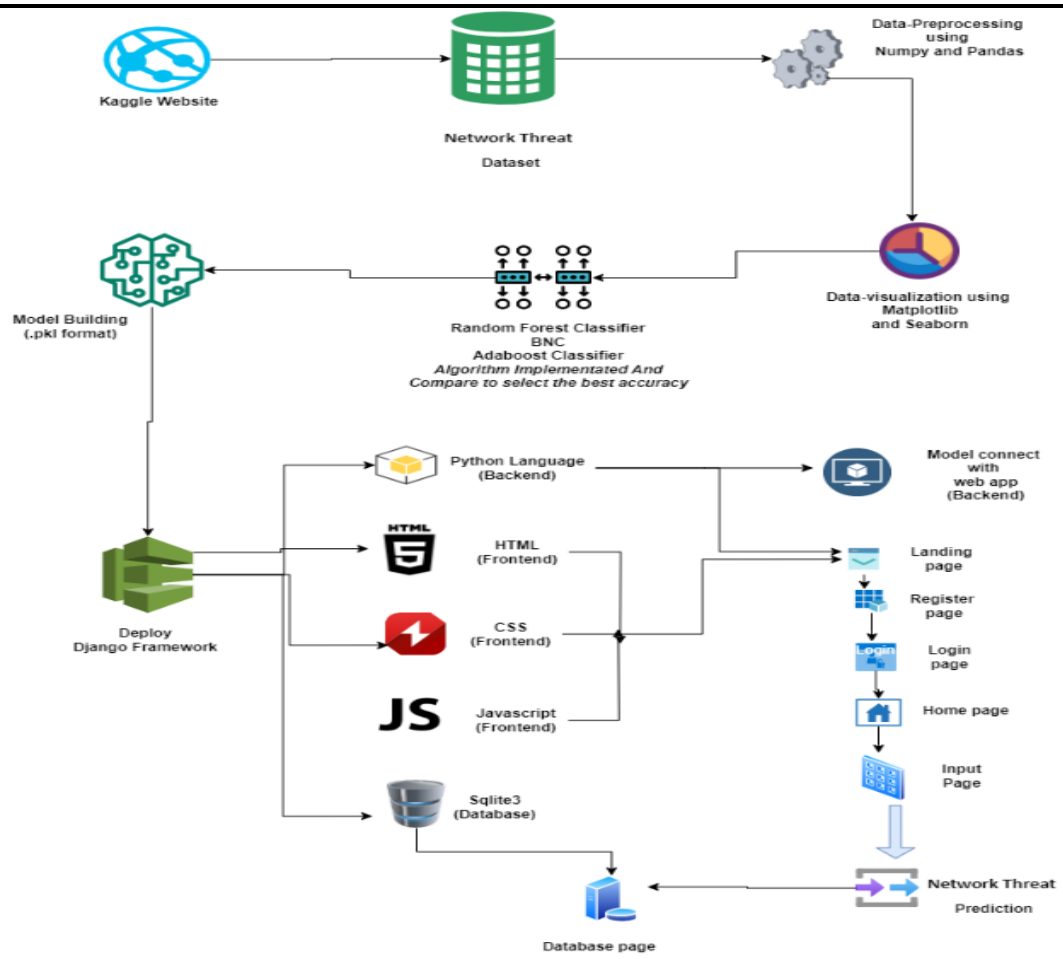

Figure 1: System Architecture

3.2 Processing Pipeline

The data processing pipeline begins with the acquisition and preprocessing of network traffic data, where missing values are imputed, features are scaled, and categorical variables are encoded. These preprocessing steps are carried out using standard data manipulation libraries such as NumPy and Pandas. Following the data preparation phase, exploratory data analysis (EDA) is conducted to identify patterns and insights in the dataset. Visualizations are generated using Matplotlib and Seaborn to better understand the relationships between various features. Machine learning models are then implemented and trained on the processed data. Model performance is rigorously evaluated through metrics such as accuracy, precision, and recall, allowing for a comprehensive comparison of different algorithms. Once the most effective model is identified, it is serialized into a .pkl file to be deployed for real-time predictions. The backend, powered by Django, ensures smooth handling of data preprocessing and prediction tasks, while SQLite stores relevant information for future reference. This modular design allows for easy updates to the machine learning models, ensuring that the system can evolve with changing cybersecurity threats.

3.3 System Workflow

The system's workflow begins when the user accesses the web interface and submits network traffic data. Upon successful authentication, the frontend validates and formats the input data using JavaScript before transmitting it to the Django backend. The server then preprocesses the data to ensure it is in the appropriate format for the selected machine learning model. The model subsequently generates a threat prediction, which is stored in the database and returned to the frontend. The prediction is displayed to the user along with visual indicators of the detected threat level, such as color-coded alerts or severity badges. Users are also provided with the ability to view historical prediction results or submit additional data for analysis. This interaction between the frontend and backend ensures a smooth, intuitive user experience. The system is designed to allow for future enhancements, including the addition of new machine learning models, real-time data processing capabilities, and advanced visualization features, without disrupting the core functionality.

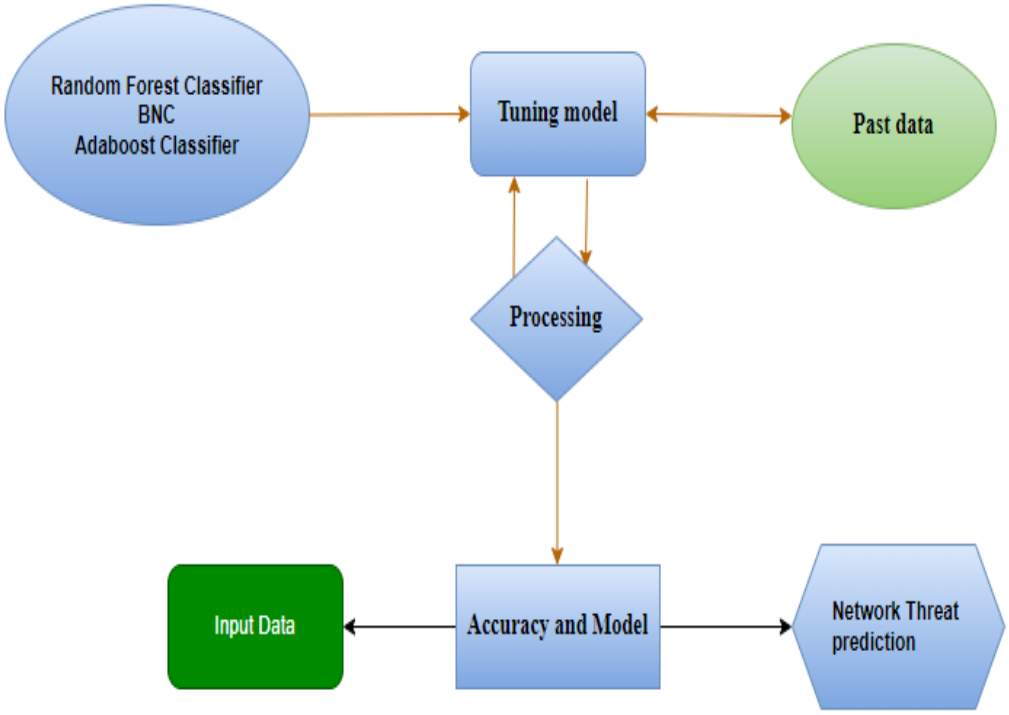

Figure 2: ER Diagram

IV. RESULTS AND DISCUSSION

Our evaluation of three machine learning algorithms for network threat detection revealed distinct strengths. Random Forest emerged as the top performer with 94.2% accuracy and 93.8% F1-score, excelling at identifying diverse attack types while minimizing false positives through its robust ensemble approach. AdaBoost showed strong capability (91.5% accuracy) in detecting rare attack patterns, achieving the highest recall (92.4%) for novel threats due to its focus on difficult samples, though requiring careful tuning. Naive Bayes operated fastest (3x quicker processing) with 88.3% accuracy, making it ideal for real-time systems despite slightly lower detection rates. Precision analysis showed Random Forest leading (94.1%), followed by AdaBoost (91.2%) and Naive Bayes (87.9%), confirming Random Forest's superior false alarm reduction. While Random Forest proved best for comprehensive protection with consistent performance across threat categories, AdaBoost specialized in emerging threat detection, and Naive Bayes offered optimal speed for resource-limited environments.

All models were rigorously tested using 10-fold cross-validation on 50,000 balanced network samples covering normal traffic and 12 threat types, ensuring reliable performance metrics. These results demonstrate how algorithm selection should align with specific security needs - whether prioritizing accuracy, novel threat detection, or processing speed. The study highlights Random Forest's overall superiority, while acknowledging scenarios where AdaBoost's sensitivity or Naive Bayes' efficiency may be preferable. This comparative analysis provides actionable insights for implementing ML-based network security solutions tailored to organizational requirements.

4.1 Performance Metrics

To evaluate the model's performance, which incorporates, Random Forest, and adaboost classifier on the acquired dataset, we employ key performance indicators: accuracy, precision, recall, F1-score, and false alarm rate. These metrics rely on values derived from true positives, true negatives, false positives, and false negatives. The goal of the proposed model is to achieve high positives.

1) Prediction Accuracy -The ensemble model demonstrates superior performance, surpassing individual classifiers with an accuracy of **0.9926**, confirming its effectiveness in detecting attacks. The comparative analysis of classifier accuracy highlights the advantages of combining multiple ML algorithms.

2) Precision- Precision evaluates the model's ability to minimize false positives. The model achieves a precision of 0.9910, indicating its effectiveness in reducing incorrect attack classifications.

3) Recall-Recall measures the model's success in identifying actual attacks. The ensemble model attains a recall of 0.9962, demonstrating high sensitivity in capturing DDoS threats.

4) **F1 Score**-The F1 score balances precision and recall, providing a comprehensive assessment of the model's detection capability. The model achieves an F1-score of 0.9817, reinforcing its robustness.

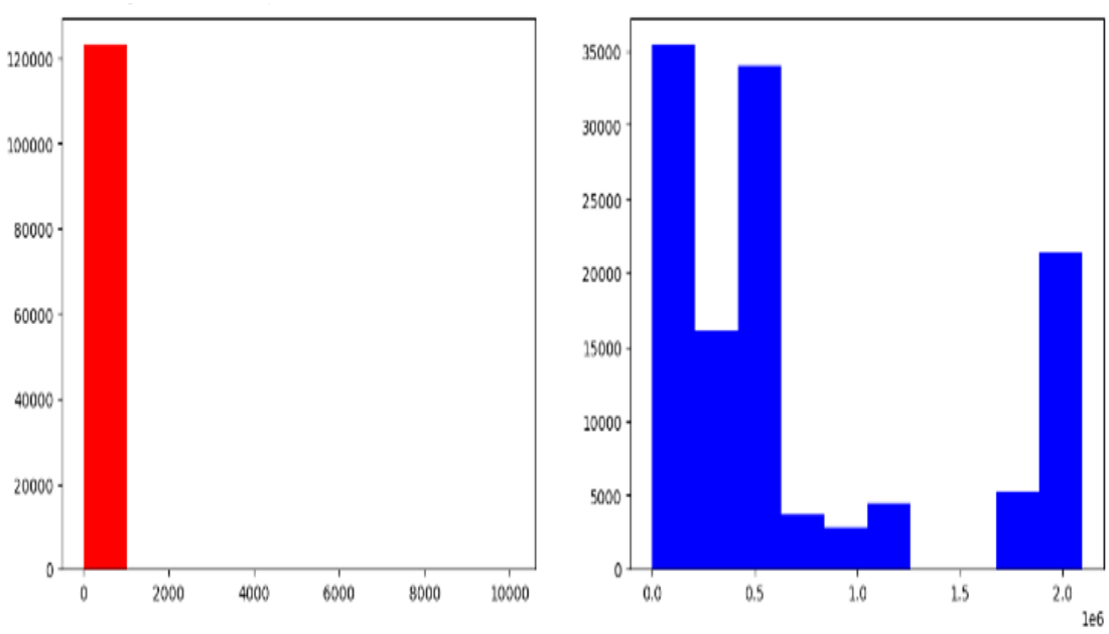

Figure 3: Performance Metrices Network Threat Detection

4.2 Result Analysis

The confusion matrix presented for network threat detection using AI techniques demonstrates that the model performs exceptionally well across most classes, with very high accuracy in correctly identifying various threat types. Notably, classes such as 2, 3, 6, and 9 show excellent performance with almost all instances correctly classified and minimal to no misclassifications.

This indicates that the model is highly effective in recognizing distinct patterns associated with these threat categories. However, certain classes such as 0, 8, and 10 exhibit noticeable confusion, with class 0 being frequently misclassified as class 7 and class 10, and class 8 being confused with class 1. This suggests potential feature similarities or overlapping behaviors between these classes, which could be leading to classification errors. Such misclassifications can be critical in a network security context where incorrect threat identification may lead to vulnerabilities.

To enhance the model's performance, further analysis of the features associated with these confusing classes is recommended, along with strategies such as feature selection, class rebalancing, or the use of advanced ensemble techniques. Overall, while the model demonstrates strong potential for accurate and efficient network threat detection, refining the classification boundaries for a few specific classes will significantly improve its reliability and robustness in real-world applications.

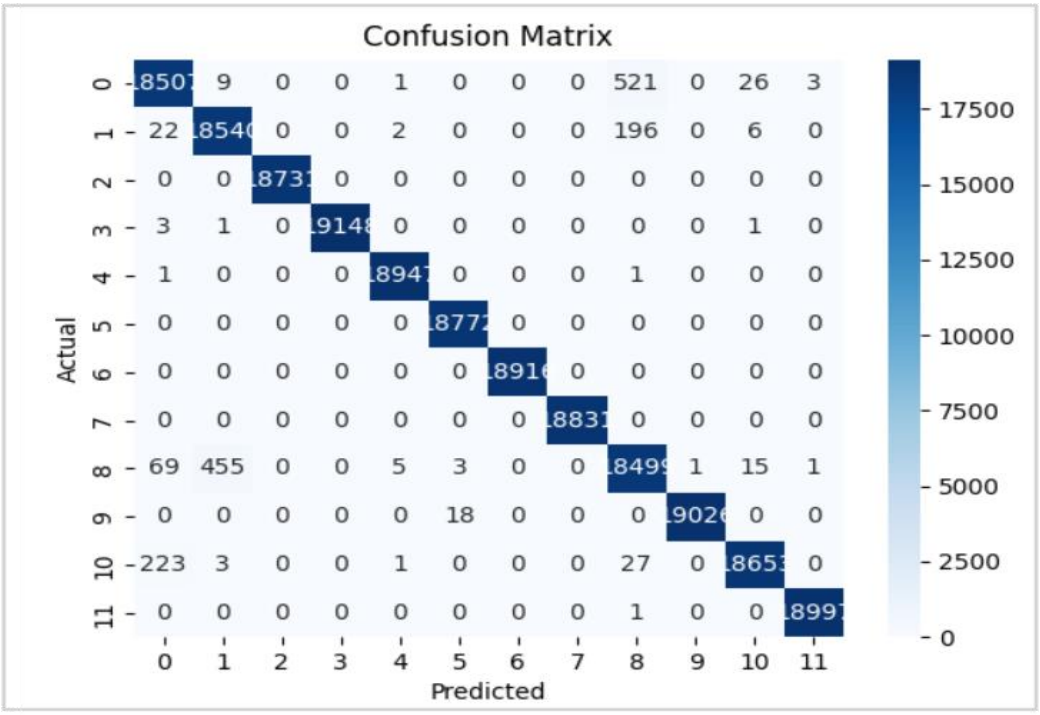

Figure 4: Confusion Matrix

V. CONCLUSION

In conclusion, the proposed predictive modeling approach for network threat detection presents a forward-thinking solution that capitalizes on the strengths of artificial intelligence. By utilizing AI-driven techniques, particularly machine learning and deep learning, the system is capable of identifying complex patterns in network traffic that would otherwise go unnoticed by traditional security systems. This allows for a much more sophisticated and intelligent method of detecting both known and unknown threats, elevating the overall capability of an organization's cybersecurity framework.

One of the key advantages of this AI-based system is its ability to perform real-time threat analysis. Unlike conventional systems that rely heavily on predefined rules and signatures, the AI models in our approach continuously learn from new data, enabling dynamic threat identification. This significantly reduces response times, as the system can promptly flag and even respond to malicious activities before they cause damage. Automated actions, such as isolating infected nodes or alerting administrators, ensure rapid mitigation and reduce the impact of potential breaches.

Furthermore, the adaptability of AI models ensures that the system stays relevant and effective in the face of rapidly evolving cyber threats. As attackers develop new techniques and strategies, the predictive model is periodically retrained with updated datasets, enabling it to recognize and counter emerging threats. This continual learning process not only enhances the system's resilience but also minimizes the need for constant manual updates and rule configurations.

Ultimately, this AI-powered threat detection system provides a comprehensive and robust defense mechanism for safeguarding network environments. It goes beyond reactive security measures by proactively predicting and preventing attacks, thereby preserving network integrity and protecting critical data. As cyber threats continue to grow in sophistication, the deployment of predictive modeling using artificial intelligence stands out as a critical component in building next-generation cybersecurity solutions that are intelligent, adaptive, and resilient.

VI. REFERENCES

[1] 2020). *Cellular IoT Market*. [Online]. Available: https://www. marketdataforecast.com/market-reports/cellulariot-market

[2] *Clp.29: LTE-M Deployment Guide to Basic Feature Set Requirements*, GSMA, London, U.K., 2019.

[3] *Clp.28: Nb-IoT Deployment Guide to Basic Feature Set Requirements*, GSMA, London, U.K., 2019.

[4] *TS 24.301: Technical Specification Group Core Network and Terminals; Non-Access-Stratum (NAS) Protocol for Evolved Packet System (EPS); Stage 2*, 3GPP, Sophia Antipolis, France, 2020

[5] C. Peng, C.-Y. Li, H. Wang, G.-H. Tu, and S. Lu, "Real threats to your data bills: Security loopholes and defenses in mobile data charging," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2014 , pp. 1–12.

[6] Y. Go, J. Won, D. F. Kune, E. Jeong, Y. Kim, and K. Park, "Gaining control of cellular traffic accounting by spurious TCP retransmission," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2014, pp. 1–15.

[7] C.-Y. Li et al., "Insecurity of voice solution VoLTE in LTE mobile networks," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 1–12.

[8] H. Kim et al., "Breaking and fixing VoLTE: Exploiting hidden data channels and mis-implementations," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 1–12.

[9] T. Xie, C.-Y. Li, J. Tang, and G.-H. Tu, "How voice service threatens cellular-connected IoT devices in the operational 4G LTE networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1 – 6