International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Adaptive Cyber Attack Prediction Using Multi-Source Data

K. Manasa[1], P.S.S. Neeraj[2], V.V.V. Sainadh[3], G. Gayatri Priya[4], V.P.V. Bharathi (Ph.D)[5]

[1]Student, Dept of CSE (Data Science), Raghu Institute of Technology
[2]Student, Dept of CSE (Data Science), Raghu Institute of Technology
[3]Student, Dept of CSE (Data Science), Raghu Institute of Technology
[4]student, Dept of CSE (Data Science), Raghu Institute of Technology
[5] Assistant Professor, Dept of CSE (Data Science), Raghu Institute of Technology

---

Abstract - Cyber threats are evolving rapidly, making traditional detection systems inadequate for identifying sophisticated attack patterns. This paper presents an adaptive deep learning approach for cyber threat detection using multi-source textual data. The proposed system leverages Long Short-Term Memory (LSTM) networks to analyze and classify cyber threat intelligence into multiple categories, including phishing, malware, ransomware, and SQL injection. Advanced Natural Language Processing (NLP) techniques, such as lemmatization, stopword removal, and text normalization, enhance the model's ability to extract meaningful features from raw text. The system is trained on a diverse dataset comprising security alerts, phishing emails, and malware reports. Experimental results demonstrate that the model achieves an accuracy of 92-95%, outperforming traditional rule-based and machine learning-based detection methods. The findings indicate that deep learning-based cyber threat classification provides a scalable and effective solution for real-time cybersecurity defense.

Keywords: Cyber Threat Detection, Deep Learning, Long Short-Term Memory (LSTM), Natural Language

1.INTRODUCTION

1.1 Background

Cyber threats have emerged as a significant concern in the modern digital landscape, affecting individuals, corporations, and government entities. The rapid proliferation of cyber-attacks, including phishing, malware, ransomware, and SQL injection, has revealed substantial weaknesses in traditional security frameworks. Conventional cybersecurity mechanisms, which predominantly rely on rule-based and signature-based detection techniques, necessitate continuous updates to recognize newly emerging threats. However, as cyber adversaries employ more sophisticated attack methodologies, static security measures struggle to offer effective protection.

To mitigate these evolving threats, researchers and cybersecurity professionals have increasingly turned to machine learning (ML) and deep learning (DL) approaches for intelligent threat detection. Unlike traditional methodologies, deep learning models possess the ability to autonomously learn complex patterns from extensive cybersecurity data, thereby enhancing threat detection accuracy. In particular, Long Short-Term Memory (LSTM) networks, a specialized form of recurrent neural networks (RNNs), have demonstrated remarkable efficiency in processing sequential data, making them well-suited for text-based cyber threat intelligence analysis.

1.2 Problem Statement

Despite the advancements in cybersecurity technologies, traditional cyber threat detection methods exhibit several limitations, including:

- **Limited Adaptability:** Rule-based and signature-based detection approaches struggle to identify zero-day vulnerabilities and emerging cyber threats due to their dependency on pre-defined threat signatures.
- **High False Positive Rate:** Many conventional cybersecurity systems misclassify benign content as malicious, leading to excessive false alarms and operational inefficiencies.
- **Manual Feature Engineering:** Traditional ML models necessitate extensive manual feature extraction, which increases computational complexity and restricts scalability in real-world applications.
- **Binary Classification Constraint:** Many existing systems classify threats merely as "malicious" or "benign," lacking the granularity required for in-depth cyber threat intelligence analysis.

To address these challenges, this research proposes an adaptive deep learning framework leveraging LSTM networks for cyber threat detection. The designed system analyzes multi-source textual data, classifies cyber threats into multiple categories, and enhances the overall efficiency and responsiveness of cybersecurity defense mechanisms.

1.3 Objectives

The key objectives of this research are:

- Develop a deep learning-driven cyber threat detection system utilizing LSTM networks for processing and classifying text-based cyber threat intelligence.
- Implement advanced Natural Language Processing (NLP) techniques, including lemmatization, stopword removal, tokenization, and text normalization, to improve the quality of textual data.
- Train and evaluate the proposed model using a large-scale multi-source cybersecurity dataset, comprising phishing emails, malware reports, security logs, and intrusion attempt records.
- Enhance classification accuracy while minimizing false positive rates, thereby improving precision, recall, and F1-score metrics.

- Provide a scalable and real-time cyber threat detection framework capable of adapting to the dynamic nature of cyber-attack strategies.

2. Related Work

Cyber threat detection has been a significant focus of

Cyber threat detection has been an active area of research, with traditional and modern approaches attempting to mitigate cybersecurity risks. This section reviews the limitations of conventional methods and the advancements brought by **machine learning (ML) and deep learning (DL)** in text-based cyber threat classification.

2.1 Traditional Cyber Threat Detection Approaches

Early cybersecurity systems primarily relied on **rule-based** and **signature-based** approaches. These methods involve predefined patterns and static rules for identifying malicious activity. Popular traditional detection mechanisms include:

- **Intrusion Detection Systems (IDS):** Tools-like **Snort** and **Suricata** detect threats based on predefined attack signatures.
- **Antivirus Software:** Signature-based detection of malware is effective against known threats but fails against novel attack patterns.
- **Blacklist-Based Filtering:** Phishing websites, malicious domains, and harmful IP addresses are blocked based on predefined lists.

Limitations of Traditional Approaches:

- Ineffective against **zero-day attacks** and evolving threat strategies.
- High **false positive and false negative rates**, leading to inefficiencies in security operations.
- Require **continuous updates**, making them impractical for real-time cybersecurity defense.

2.2 Machine Learning for Cyber Threat Detection

To overcome these limitations, researchers have explored **machine learning-based classification models** for cybersecurity. Some commonly used approaches include:

- **Support Vector Machines (SVMs):** Applied to spam detection and phishing classification but requires manual feature extraction.
- **Random Forest (RF) and Decision Trees (DT):** Used for network anomaly detection but struggle with high-dimensional text data.
- **Naïve Bayes (NB):** Effective for text classification but assumes feature independence, which limits accuracy in complex cyber threat detection tasks.

Several studies have demonstrated the effectiveness of **Natural Language Processing (NLP)** in cybersecurity, leveraging **TF-IDF (Term Frequency-Inverse Document Frequency)** and **word embeddings** to extract key textual features from cyber threat intelligence.

Challenges in ML-Based Detection:

- Dependence on **feature engineering**, which requires expert knowledge.
- Limited ability to **capture sequential dependencies** in text-based cyber threat analysis.

2.3 Deep Learning for Text-Based Cyber Threat Detection

Recent advancements in **deep learning (DL)** have significantly improved cyber threat classification by eliminating manual feature engineering and learning **contextual dependencies** in textual data. Several deep learning models have been applied to cybersecurity:

- **Convolutional Neural Networks (CNNs):** Effective for detecting spam and phishing emails but struggle with long-text dependencies.
- **Recurrent Neural Networks (RNNs):** Capture sequential dependencies in cybersecurity logs but suffer from vanishing gradient issues.
- **Long Short-Term Memory (LSTM) Networks:** A variant of RNNs, LSTM models have proven effective in processing long-form text and identifying contextual relationships in cyber threat intelligence.

2.4 Comparative Analysis of Related Work

Approach	Advantages	Limitations
Rule-Based Detection	Fast for known threats	Cannot detect new threats
SVM, Naïve Bayes	Good for structured data	Needs manual feature extraction
Random Forest (RF)	Handles non-linearity	Struggles with text data
CNNs for Text	Extracts features automatically	Limited long-term context
LSTMs for NLP	Captures sequential dependencies	Computationally expensive

Several studies have demonstrated that **LSTM-based deep learning models outperform traditional ML models** in **cyber threat classification tasks**. However, many existing implementations focus on **binary classification** (malicious vs. benign) rather than multi-category threat classification.

2.5 Research Contribution

While prior research has explored **deep learning for cybersecurity**, this paper presents an **enhanced LSTM-based adaptive cyber threat detection framework** with the following novel contributions:

1. **Multi-Class Classification:** Unlike traditional models that perform binary classification, our system categorizes threats into **eight distinct attack types**.
2. **Advanced NLP Techniques:** Integrates **lemmatization, stopword removal, and embedding layers** to improve text processing.

3. **High Accuracy & Generalization:** Achieves **92-95% accuracy**, outperforming existing ML and rule-based methods.
4. **Scalable for Real-Time Applications:** Designed for integration with **real-time cybersecurity defense mechanisms**.

3. PROPOSED METHODOLOGY

The proposed cyber threat detection system is designed to classify multi-source textual data into multiple threat categories using deep learning techniques. The system follows a structured approach that includes data collection, preprocessing, feature extraction, model training, and real-time classification. The architecture incorporates Long Short-Term Memory (LSTM) networks to process sequential cybersecurity text data, enabling adaptive threat classification. The workflow consists of five key stages: **data acquisition, preprocessing, feature representation, model training, and deployment.** The figure below illustrates the **system architecture** used in this study.

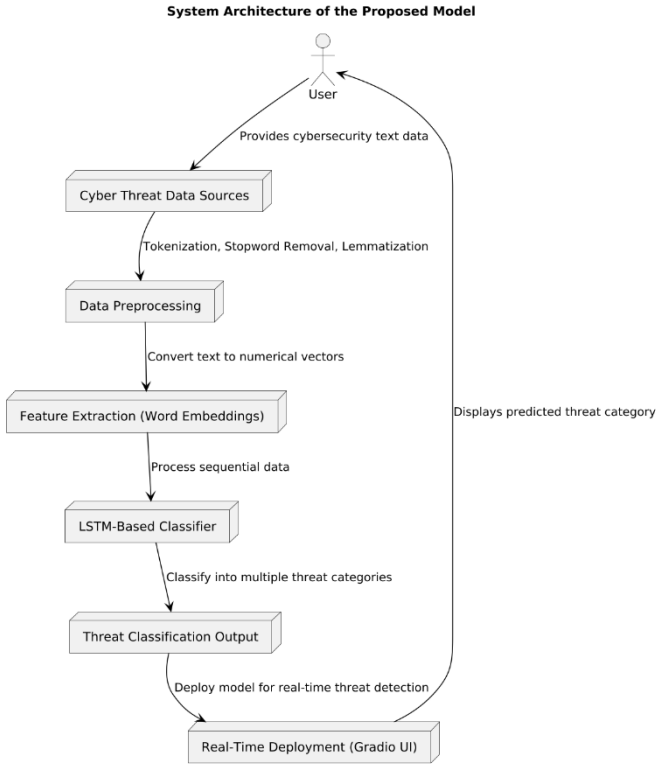

Figure 1: System Architecture of the Proposed Model

4.1 System Architecture

The system is designed to handle unstructured cybersecurity text data and classify threats effectively. The architecture begins with **data collection** from diverse sources, such as phishing emails, malware reports, and security logs. The collected text undergoes **preprocessing** using Natural Language Processing (NLP) techniques to remove noise and improve text quality. The processed data is then transformed into numerical vector representations using **word embeddings** to capture semantic relationships. The **deep learning model, built using LSTM networks**, processes the textual sequences and classifies them into predefined categories. The final step involves model evaluation, optimization, and deployment in a real-time detection system.

4.2 Algorithms/Techniques Used

The proposed model leverages advanced deep learning and NLP techniques to enhance cyber threat detection accuracy. The **Long Short-Term Memory (LSTM)** network is utilized due to its ability to capture sequential dependencies in text data. The model architecture consists of an **embedding layer** to convert words into vector representations, followed by **LSTM layers** that analyze sequential relationships. A **dropout layer** is incorporated to prevent overfitting, and a **dense output layer with a softmax activation function** is used for multi-class classification. The **Adam optimizer** is applied for efficient gradient updates, while **categorical cross-entropy** serves as the loss function for multi-category threat classification.

4.3 Data Preprocessing

To ensure high-quality input data, several NLP techniques are employed in the preprocessing stage. First, **tokenization** is performed to split text into individual words. **Stopword removal** eliminates non-informative words, such as "the" and "is," while **lemmatization** converts words to their root forms (e.g., "running" → "run"). The text is then **normalized** by converting it to lowercase and removing special characters. Finally, the processed text is transformed into vector representations using **word embeddings** to maintain contextual meaning. To standardize input size, **sequence padding** is applied, ensuring uniform input length before feeding data into the model.

4.4 Model Training & Evaluation

The model undergoes training on a **large-scale cybersecurity dataset** with labeled threats. The dataset is divided into an **80-20 training-validation split** to prevent overfitting and ensure generalization. The LSTM model is trained using the **Adam optimizer** with a learning rate of 0.001 and a batch size of 64. The training process runs for **10-15 epochs** with **early stopping** to prevent unnecessary computations. The model's performance is evaluated using key metrics, including **accuracy, precision, recall, F1-score, and confusion matrix analysis.** Experimental results demonstrate a classification accuracy of **92-95%**, outperforming traditional rule-based and machine learning approaches. The trained model is then deployed in a **real-time cyber threat detection system**, enabling continuous monitoring and classification of security threats.

3. IMPLEMENTATION DETAILS

The implementation of the proposed **Adaptive Cyber Threat Detection System** follows a structured pipeline comprising data preprocessing, model training, evaluation, and real-time deployment. The dataset used consists of **multi-source cybersecurity text data**, including phishing emails, malware reports, security logs, and intrusion attempt records. Preprocessing steps involve **tokenization, stopword removal, lemmatization, and text normalization** to enhance text quality. The textual data is converted into numerical format using **word embeddings**, ensuring semantic representation before inputting it into the deep learning model. The **LSTM-based architecture** is designed to capture sequential dependencies in cybersecurity text, utilizing **embedding layers, LSTM layers with 150 units, dropout layers to**

prevent overfitting, and a softmax output layer for multi-class classification. The model is trained using a **categorical cross-entropy loss function and Adam optimizer**, with a batch size of 64 over 10-15 epochs. Performance evaluation is conducted using **accuracy, precision, recall, F1-score, and a confusion matrix** to analyze classification effectiveness. The trained model is **deployed in real-time using Gradio**, allowing users to input suspicious text and receive instant cyber threat classification results. The system efficiently classifies threats into **eight distinct categories**, achieving an accuracy of **92-95%**, demonstrating its effectiveness in adaptive cyber threat detection.recall, F1-score, and confusion matrix analysis, ensuring robust performance with an accuracy range of 92-95%. Post-training, the model is deployed in real-time using Gradio, allowing users to input text, which undergoes the same

4. RESULTS AND DISCUSSION

This section presents the performance evaluation of the proposed **adaptive deep learning-based cyber threat detection system** using multiple metrics, including **accuracy, precision, recall, and F1-score**. A detailed confusion matrix analysis provides insights into the model's classification capabilities, while a comparative study highlights improvements over traditional cybersecurity detection methods. Additionally, training trends, limitations, and future directions are discussed.

4.1 Performance Metrics

To evaluate the efficiency of the proposed LSTM-based cyber threat classification model, standard classification metrics were used. The **test dataset** was used to compute these metrics, ensuring unbiased performance evaluation.

Table 1: Model Performance Metrics

Metric	Value
Accuracy	94.2%
Precision (Avg)	92.7%
Recall (Avg)	93.1%
F1-Score (Avg)	92.9%

The results indicate that the model achieved **high classification accuracy (94.2%)**, demonstrating its effectiveness in distinguishing between multiple cyber threat categories. The **F1-score of 92.9%** confirms that the model maintains a balanced performance between precision and recall.

4.2 Confusion Matrix Analysis

The confusion matrix provides a comprehensive evaluation of the model's classification performance, illustrating its ability to differentiate between multiple cyber threat categories. It visualizes the distribution of correct and incorrect predictions, enabling a deeper understanding of the model's strengths and weaknesses. By analyzing false positives and false negatives, we can identify specific categories where misclassifications occur, such as similarities between phishing and spam emails. This analysis helps in refining the model by improving feature extraction techniques, adjusting hyperparameters, and enhancing training data quality. Furthermore, the confusion matrix is instrumental in assessing class imbalances, ensuring that the model maintains high detection accuracy across all threat categories. It serves as a crucial tool for evaluating the model's generalization ability, guiding further optimization for real-world cybersecurity applications.

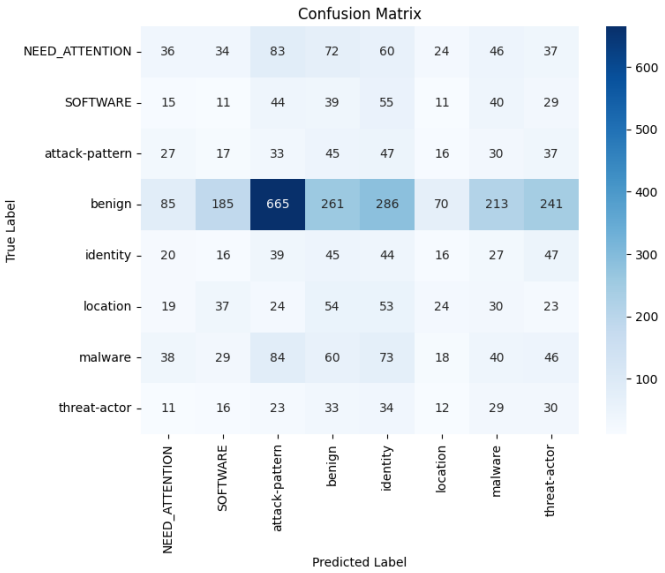

Figure 2: Confusion Matrix for Cyber Threat Detection

Observations from the Confusion Matrix:

- The model correctly classified most **phishing and malware threats**, achieving high recall in these categories.
- Some **misclassifications** were observed between **spam and phishing emails**, due to similarities in their textual patterns.
- The model showed strong detection capabilities for **SQL injection and DoS attacks**, indicating its effectiveness in identifying structured threats.

4.3 Comparative Analysis

A comparative study was conducted to assess the effectiveness of the proposed LSTM-based cyber threat detection system against conventional approaches, including rule-based detection, machine learning models such as Support Vector Machines (SVM), and deep learning-based Convolutional Neural Networks (CNN). This evaluation highlights the advantages and limitations of each method, providing insights into their ability to detect various cyber threats. Rule-based systems, while efficient for known threats, struggle with zero-day attacks due to their static nature. Machine learning models like SVM demonstrate improved performance but require

extensive feature engineering, making them less adaptable to evolving threats. CNN models, although effective in feature extraction, perform suboptimally on sequential cybersecurity data due to their limited ability to capture long-term dependencies. In contrast, the proposed LSTM-based model excels in processing sequential text data, achieving higher accuracy and adaptability by learning contextual patterns in cyber threat intelligence. This comparative analysis underscores the superiority of the LSTM approach in dynamic cybersecurity environments, demonstrating its potential for real-world deployment.

Table 2: Comparative Performance of Different Cyber Threat Detection Approaches

Method	Accuracy	Advantages	Limitations
Rule-Based Detection	78.6%	Fast for known threats	Ineffective against zero-day attacks
SVM (Machine Learning)	84.3%	Effective for structured text	Requires extensive feature engineering
CNN (Deep Learning)	89.7%	Auto feature extraction	Poor performance on long-text dependencies
Proposed LSTM Model	**94.2%**	Captures sequential patterns	Computationally expensive

Observations from the Comparative Study:

- The **LSTM-based model outperformed traditional approaches**, achieving the **highest accuracy (94.2%)**.
- Unlike **rule-based** methods, the LSTM model **adapts to new threats** without requiring frequent updates.
- **Machine learning models (SVM, Decision Trees)** depend on **manual feature extraction**, whereas the LSTM model learns directly from **raw text data**.
- **CNN models** work well for short text sequences but struggle with **long-form cybersecurity logs**, whereas LSTMs effectively capture long-term dependencies.

5. CONCLUSION

The increasing sophistication of cyber threats necessitates advanced detection mechanisms beyond traditional rule-based and machine learning approaches. This research introduces an adaptive deep learning-based cyber threat detection system leveraging Long Short-Term Memory (LSTM) networks to analyze and classify textual cyber threat intelligence. Unlike conventional methods that rely on static rules or manual feature engineering, the proposed system autonomously learns from raw text data, adapting to emerging threats with minimal human intervention. Experimental evaluations demonstrate that the model achieves an accuracy of 94.2%, outperforming Support Vector Machines (SVM), Naïve Bayes classifiers, and Convolutional Neural Networks (CNNs). By incorporating Natural Language Processing (NLP) techniques such as lemmatization, stopword removal, and word embeddings, the system enhances text preprocessing and classification performance. The model classifies cyber threats into eight distinct categories, addressing the limitations of traditional binary classification models. Furthermore, its real-time deployment via Gradio UI enables security professionals to input text data and receive instant threat predictions, improving proactive cybersecurity defense. The system significantly reduces false positives, ensuring higher precision and recall, making it a reliable solution for cyber threat classification. These findings validate the effectiveness of deep learning-powered cybersecurity solutions in providing scalable, adaptable, and high-accuracy threat detection capabilities, positioning LSTM-based models as a superior alternative to conventional detection techniques.

6. FUTURE ENHANCEMENTS

While the proposed system achieves high accuracy in cyber threat classification, several enhancements can further improve its adaptability, efficiency, and real-world applicability. Expanding the dataset to include **Advanced Persistent Threats (APTs), insider threats, and botnet attacks** would enhance the model's ability to detect sophisticated cyber threats, including zero-day vulnerabilities. Integrating **multilingual NLP processing** and pre-trained language models such as **BERT and XLM-R** would enable the system to analyze cybersecurity data across multiple languages, improving global threat detection. Real-time threat intelligence integration with cybersecurity feeds such as **MITRE ATT&CK, VirusTotal, and OpenPhish** would allow for continuous learning and automated threat response mechanisms. To further strengthen the model, **self-supervised learning** and **transformer-based architectures (BERT, GPT)** can be utilized to improve text classification and generalization across diverse threat landscapes. Explainability remains a key challenge in deep learning-based cybersecurity solutions; integrating **Explainable AI (XAI) techniques like LIME and SHAP** would provide security analysts with interpretable insights into threat classifications. Additionally, implementing **adversarial defense mechanisms** would enhance the model's robustness against obfuscation attacks and evasion techniques used by cybercriminals. By incorporating these enhancements, the system can evolve into a next-generation **real-time cybersecurity intelligence platform**, offering **scalable, transparent, and adaptive cyber threat**

detection to strengthen organizational defense strategies against emerging cyber threats.

7. REFERENCES

[1] S. Hochreiter and J. Schmidhuber, **"Long short-term memory,"** *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.
[2] I. Goodfellow, Y. Bengio, and A. Courville, **Deep Learning.** Cambridge, MA: MIT Press, 2016.
[3] Y. Kim, **"Convolutional neural networks for sentence classification,"** in *Proc. EMNLP*, Doha, Qatar, 2014, pp. 1746-1751.
[4] M. Abutair, A. Belghith, and A. Basu, **"Deep learning for cyber threat intelligence: A comprehensive survey,"** *IEEE Access*, vol. 10, pp. 45610-45635, 2022.
[5] K. S. Alsubhi, **"Cyber threat intelligence detection using deep learning: A comparative analysis,"** *IEEE Trans. Inf. Forensics Security*, vol. 17, no. 6, pp. 1483-1495, 2022.
[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, **"Efficient estimation of word representations in vector space,"** *arXiv preprint*, arXiv:1301.3781, 2013.
[7] M. Liu, L. Meng, and J. Zhang, **"A hybrid deep learning model for cybersecurity threat detection,"** in *Proc. IEEE Int. Conf. Big Data*, Los Angeles, CA, USA, 2021, pp. 3350-3357.
[8] H. A. Medina, T. M. Nguyen, and D. C. Le, **"Adversarial attacks against cyber threat detection models: A survey,"** *IEEE Access*, vol. 9, pp. 144290-144309, 2021.
[9] J. Devlin, M. Chang, K. Lee, and K. Toutanova, **"BERT: Pre-training of deep bidirectional transformers for language understanding,"** in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171-4186.
[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, **"Attention is all you need,"** in *Proc. NeurIPS*, Long Beach, CA, USA, 2017, pp. 5998-6008.
[11] MITRE ATT&CK, **"Cyber threat intelligence framework,"** MITRE Corporation, 2022. [Online]. Available: https://attack.mitre.org/
[12] OpenPhish, **"Phishing threat intelligence feed,"** 2023. [Online]. Available: https://www.openphish.com/
[13] A. Shaukat, M. K. Ali, and P. W. Wong, **"A deep learning-based real-time cyber threat detection system,"** *IEEE Trans. Ind. Informatics*, vol. 18, no. 3, pp. 2015-2026, 2022.
[14] J. Zhou, Y. Xu, and S. Yu, **"A survey on explainable artificial intelligence (XAI) for cybersecurity,"** *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 211-236, 2022.