International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Proactive Cybersecurity Using Machine Learning: Threat Detection, Diagnosis, and Propagation Analysis

Deeksha Karthik[1], Anand Zutshi[2]

[1]Oberoi International School, Grade 11, India
Email: *deekshakarthik09[at]gmail.com*

[2]Netaji Subhas University of Technology, Delhi
Email: *123anandji[at]gmail.com*

Abstract: ***Cyber threats have been rising at an increasing pace, and there is a growth in demand for their detection, classification and mitigation. The following study employs machine learning technology and a data-driven approach to analyse and combat cyber threats using a dataset which comprises network traffic data, textual context and entity relationships. We performed exploratory data analysis (EDA) to identify threat patterns, their characteristics, and explore mitigation strategies to combat digital threats. The threats have been classified based on textual data to predict diagnoses and solutions. Furthermore, network and relationship analysis have been used to identify the communication pathways, entities, and understand the propagation of cyber threats visually using graphs. Our results show that combining text analysis, graph models, and machine learning improves the accuracy of cyber threat detection, diagnosis, and mitigation. This research contributes to the field of cybersecurity by offering a framework to strengthen proactive defences against evolving cyber threats.***

Keywords: Cybersecurity, Machine Learning, Threat Classification, Graph-Based Analysis, Network Propagation

1. Introduction

The increasing pace of technological advancements in the current digital era has led to a focus on cyber threat detection for global organisations. Consequently, there is a growing demand for effective detection mechanisms, prompting a critical reassessment of existing systems. This shift necessitates a reassessment of traditional methods, which are time and resource-intensive, and majorly focus on reactive measures instead of precautionary stances (Fakiha, 2023; Sarker, 2022). The integration of machine learning and artificial intelligence is a driver for the transition from reactive to proactive threat mitigation strategies (Camacho, 2024). Large datasets can be reviewed and analysed by machine learning models in order to effectively detect anomalies that indicate attacks, therefore improving security by focusing on real-time insights (Sanmorino, 2024).

While artificial intelligence is an innovative and effective approach to threat mitigation, challenges emerge in model bias, false detections, adversarial attacks and ethical concerns (Mohammed,2024). Moreover, issues such as data privacy, scalability, and complexity of model training can hinder the functioning of machine learning models in practical situations (Rehan, 2024).

In response to these limitations, studies have proposed hybrid approaches, such as combining statistical and semantic analysis or using natural language processing to ensure threat detection is adaptable and more precise. (Shelke & Hämäläinen, 2024; Marinho & Filho, 2023).

This study explores the application of machine learning and data science to analyse cyber threats using a dataset comprising network traffic data, textual descriptions, and entity relationships.

2. Related Work

Existing research on cybersecurity primarily consists of institutional strategies for threat prevention and graph-based approaches for threat detection. Cheng and Wang (2022) propose an eight-pillar framework for higher education institutions to build a cybersecurity strategy in order to address vulnerabilities such as decentralisation and open access. Abrahams et al. (2024) provide an analysis of various cybersecurity strategies by tracing their evolution in modern organisational contexts. Their work highlights the need to integrate AI and ML technology to automate complex tasks and employ self-learning algorithms to adapt to evolving threats. Wagner and Mapp (2024) extend the focus on institutions by identifying key leadership and technical competencies required by executives to build a strategic cybersecurity program.

Jamnik et al. (2024) present a threat intelligence system using graph theory and machine learning for cyber threat detection and anomaly identification. Their framework effectively models network entities to detect unusual patterns in real time. Gulbay and Demirci (2024) analyse Advanced Persistent Threats (APTs) by encoding entities and relationships into a graph structure and performing centrality, community, and similarity analysis to identify threats. Piplai et al. (2020) develop a system that processes After Action Reports to create a fused cybersecurity knowledge graph in order to identify patterns in malware attacks. This helps analysts trace links between separate incidents and uncover hidden connections. Watts et al. detect ransomware by introducing a quantum graph-spectral fingerprinting

framework that converts the ransomware code into graphs and analyses the structure. Their model identifies disguised or modified ransomware, which often evades traditional security measures that fail against fileless and polymorphic threats.

In conclusion, past work shows that tackling cyber threats requires both strong organisational planning and adaptive technical tools.

3. Dataset Description

The study uses a dataset sourced from Kaggle (Cyber Threat Intelligence Dataset). The dataset consists of records of cybersecurity incidents and offers insights into their interconnections. The primary goal of analysing this dataset is to enhance security defences through machine learning and network analysis techniques.

This dataset includes various cyber threats such as phishing, malware, ransomware, DDoS attacks, and data breaches. Phishing refers to attempts to access sensitive information through fraudulent communication, such as emails or websites. Malware is a broad term for harmful software like viruses, worms, and Trojan horses. Ransomware locks or encrypts a user's files or entire system to demand a ransom for access. DDoS (Distributed Denial of Service) attacks overwhelm a system with traffic to disrupt normal functions. Data breaches occur when unauthorised parties gain access to confidential information, often leading to data exposure or theft.

The dataset contains key attributes to provide information for each cyber threat, including '*id', 'text', 'entries', 'relations', 'diagnosis'* and *'solution'*.

The **id** attribute is a unique identifier for each entry(e.g. 45825)

The **text** attribute is a description of the cyber threat incident, used for text-based analysis. (e.g. "*Once executed by the user the first stage malware downloads and executes the ransomware from a fixed hardcoded server list.*")

The **entities** attribute includes the information related to the threat, such as IP addresses, domains, and malware names. This data can be used to understand the main components involved in the cyber threats (e.g. {id': 44577, 'label': 'threat-actor', 'start_offset': 53, 'end_offset': 67}).

The **relations** attribute describes how various components are interconnected. (e.g. {'id': 1, 'from_id': 44599, 'to_id': 44600, 'type': 'uses'})

The **label** attribute identifies the type of cyber threat associated with the incident. (e.g. *"Malware"*)

The data was pre-processed for accurate analysis. First, textual data was normalised by converting all text to lowercase to ensure consistency. Special characters, numbers, and stopwords were removed, followed by the application of tokenisation techniques to reduce words to their root forms, such as converting "running" to "run," allowing for uniform interpretation. Furthermore, entities and their corresponding relationships were extracted from the text to support a more detailed and structured analysis.

4. Methodology

The study employs a multi-stage approach that integrates text processing, modelling networks, and machine learning classification models. Raw textual threat descriptions are processed and tokenised, with entities and their relationships extracted. The text is transformed into sentence embeddings using the Universal Sentence Encoder, and t-SNE is utilised for dimensionality reduction to facilitate visual analysis of the patterns. Threat propagation patterns and communication pathways are visualised, and a graph-based analysis is performed. For classification, LightGBM and XGBoost are trained on the embeddings and structured features. Model performance is evaluated using precision, recall, and F1-score. All experiments are conducted in Python employing established machine learning and network analysis libraries.

Exploratory Data Analysis (EDA)

The most **frequent bigrams** [Fig. 1] in the dataset include "biopass rat," "rat loader," "credential phishing," and "remote code," revealing common malware references and attack techniques. Biopass RAT is a Remote Access Trojan (RAT) that allows attackers to gain unauthorised access to a system, often used for data exfiltration and surveillance. Rat loader refers to a malware loader that installs and executes RATs. Credential phishing is an attack targeting user credentials through deceptive methods such as fake login pages and phishing emails. Remote code is linked to Remote Code Execution (RCE), a vulnerability that allows attackers to execute harmful commands on a victim's device without physical access.

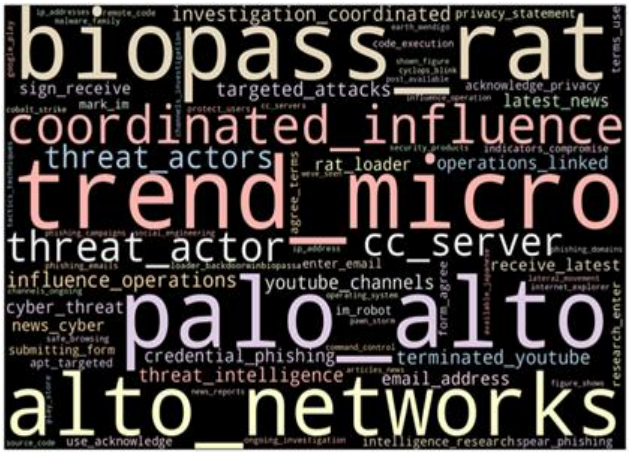

Figure 1: Bigram word cloud distribution

The most **frequent trigrams** [Fig. 2] are "biopass rat loader," "remote code execution," and "code execution vulnerability." The term "code execution vulnerability" reinforces the significance of such exploits in cybersecurity threats. Additionally, there are also mentions of "palo alto networks", thus suggesting that the cybersecurity reports may originate from known organizations. These patterns suggest a focus on malware propagation, phishing credentials, and remote code execution exploits.

Figure 2: Trigram word cloud distribution

When we compared the different types of threats, the most prominent among them was malware as can be seen in Fig 3.

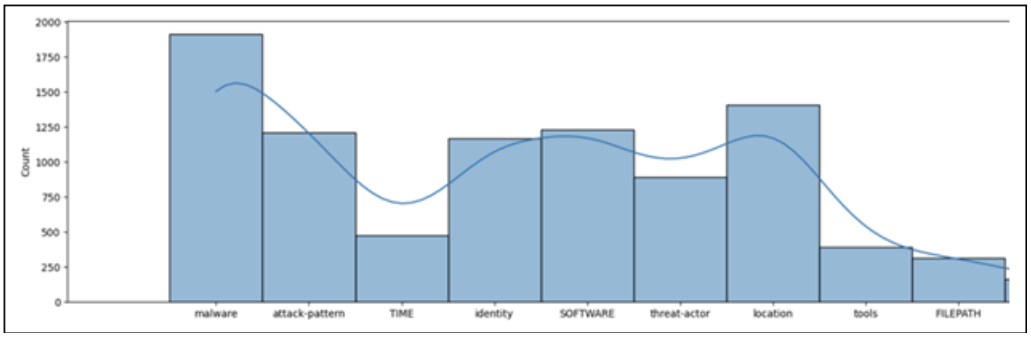

Figure 3: Threat type distribution

The Entity-Relationship graph for the dataset [Fig. 4], showcased the relation between the senders and receivers in the network. It showcased that almost all the senders were located in the middle of the network, whereas the receivers were present on the outskirts of the network.

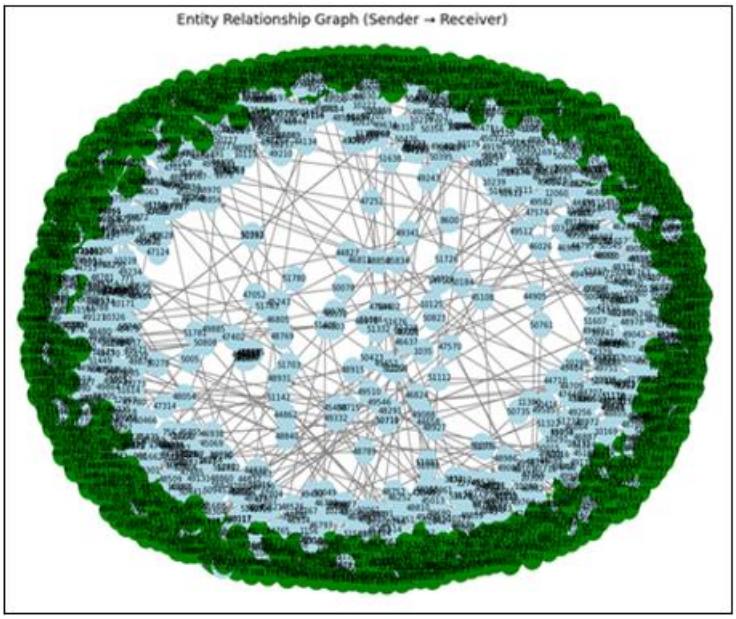

Figure 4: Entity Relationship graph between senders (blue), and receivers (green)

We further applied Louvain algorithm to determine the communities in the network [Fig. 5]. However, the only insights we got was regarding the receivers on the edges of the network. They seemed to be forming a community of types. The senders in the middle, did not form a major community.

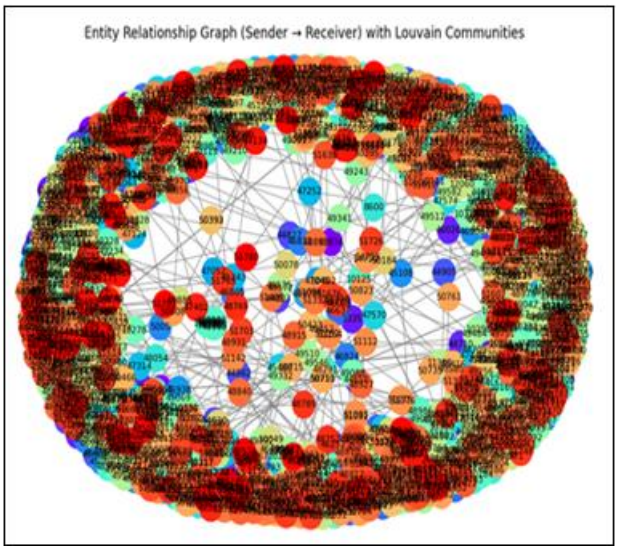

Figure 5: Louvain Communities distribution

We also performed anomaly detection in the network [Fig. 6] to determine, isolated nodes and suspicious edges. As per the observation, most of the edges stemming from the senders (in the middle) to the receivers (on the edge) of the network were suspicious; along with many senders as outliers for the network.

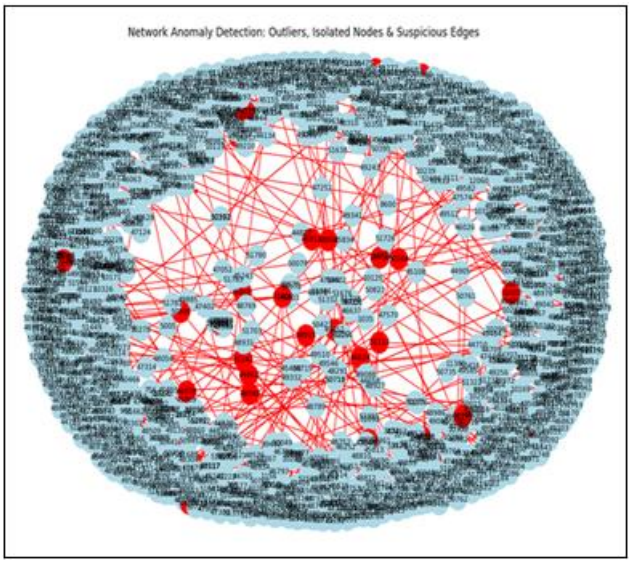

Figure 6: Anomaly detection with outlier nodes (red) and suspicious edges (red)

Based on the anomaly detection and the suspicious edges, we tried visualising the different attack patterns with different attacked / impacted nodes in the network. These patterns can be seen in Fig. 7, which showcase how the pattern evolves as the number of impacted nodes increase.

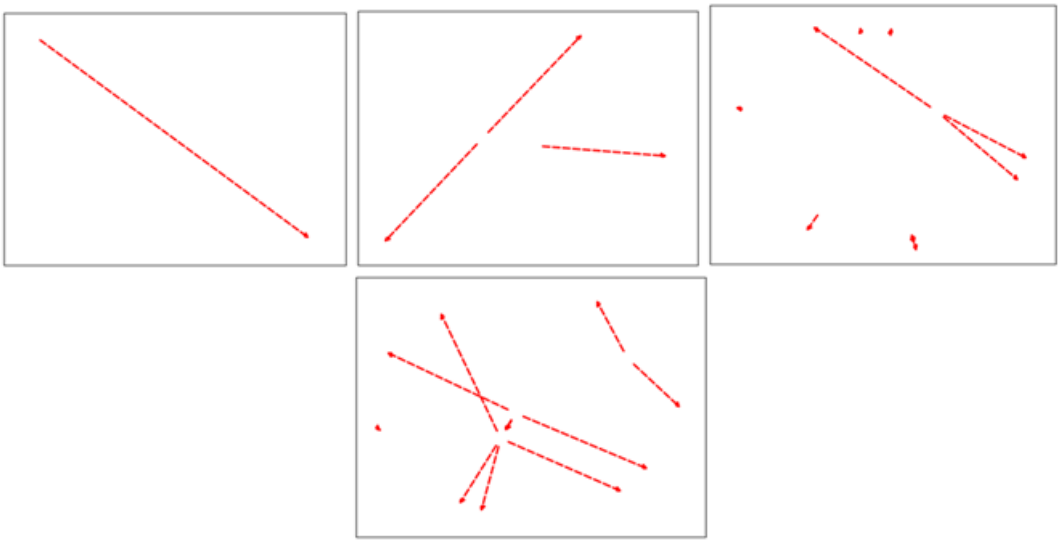

Figure 7: Attack pattern visualisations with increasing number of attackers (1,3,5,7 in clockwise from top-left)

We used the Universal Sentence Encoder of Tensorflow to generate embeddings from the text. We used TSNE to plot the embeddings on a 2-D plot and visualise the patterns and similarity. In order to ensure more relevance, we captured the size of the sentences with the size of the embeddings. We can witness longer sentences being clubbed together in the embeddings towards the left in Fig. 8.

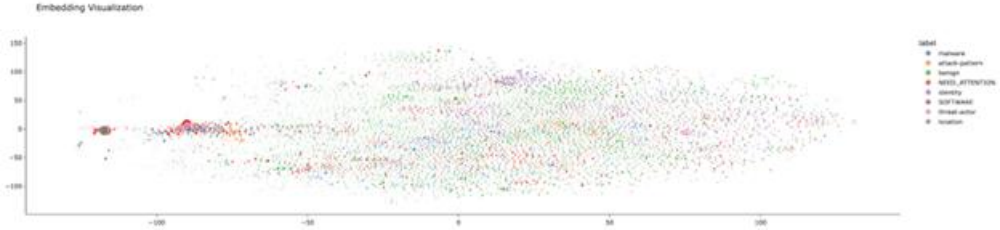

Figure 8: Embedding visualisation capturing the size / length of the sentences

Training models and metrics

Based on our analysis, we focused on training LightGBM (G. Ke et al, 2017) and XGBoost (T. Chen and C. Guestrin, 2016) specifically to classify the type of threats. We specifically picked these two models because both are gradient boosting frameworks that are highly efficient and well-suited for handling structured/tabular data. XGBoost is known for its robustness and regularization capabilities, while LightGBM is optimized for speed and memory efficiency, especially when working with large datasets. LightGBM also uses histogram-based algorithms which can lead to faster training.

While training these models, we looked at precision, recall, F1 score, as well as feature importances. These metrics were crucial to understanding not just overall accuracy but the ability of our models to correctly identify true threat categories. Precision measures the proportion of correctly identified threats among all predicted threats, recall evaluates how well the model detects all actual threats, and the F1 score provides a balance between precision and recall. Feature importance helped us identify which variables contributed most to the classification decisions (A. Géron, 2019).

Our training incorporated K-Fold cross-validation using a fold of 5, during which we ensured to pick the best model with the lowest loss and highest accuracy (T. Hastie, R. Tibshirani, and J. Friedman, 2009).

In order to understand the model and perform model interpretation, we performed key steps. Firstly, we highlighted key features which the model identified as important. Secondly, we visualised the decision tree for the model. Lastly, we combined the embeddings and the feature importances to perform model interpretation on the input sentences, understanding the model.

5. Results and Discussion

Out of the 512 features generated from the sentence embeddings, our LightGBM model gave score to the features. Fig 9 showcases the features which have a score of more than 80. This threshold was decided after considering the feature score distribution across all the 512 features. The distribution of the feature scores can be seen in Fig 10.

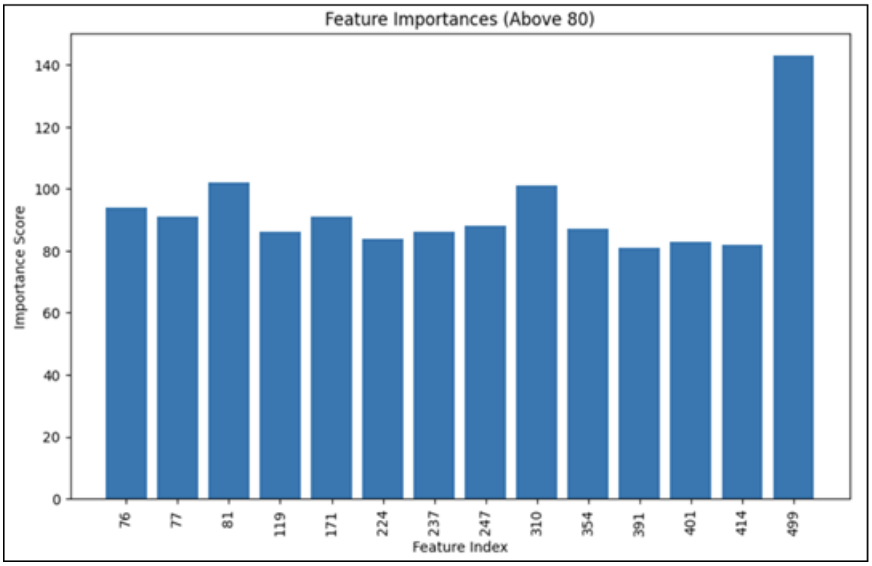

Figure 9: Feature scores of features having score > 80

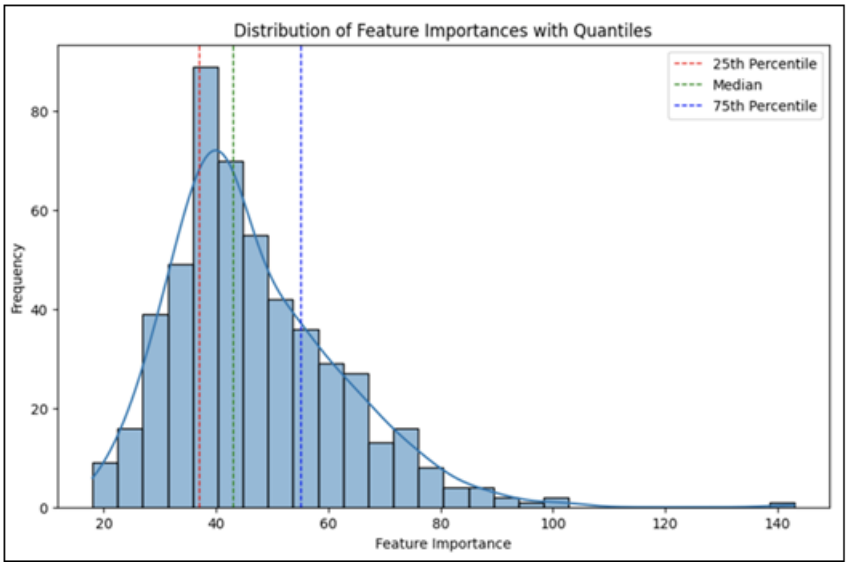

Figure 10: Feature score distribution for LightGBM

Fig. 11 showcases the different scores for the threats, including their precision, recall and f1 scores.

	precision	recall	f1-score	support
NEED_ATTENTION	0.81	0.82	0.81	392
SOFTWARE	0.91	0.64	0.75	244
attack-pattern	0.89	0.71	0.79	252
benign	0.85	0.97	0.91	2006
identity	0.89	0.70	0.78	254
location	0.85	0.72	0.78	264
malware	0.90	0.80	0.85	388
threat-actor	0.94	0.70	0.80	188
accuracy			0.86	3988
macro avg	0.88	0.76	0.81	3988
weighted avg	0.86	0.86	0.85	3988

Figure 11: LightGBM scores for different types of threats

We did a reverse embedding to highlight the regions of importances considering the top features and the different types of threats. Fig 12 highlights the model's interpretation when detecting different types of threats from the input sentence / description.

Figure 12: LightGBM model interpretation for different threat descriptions

Fig. 13 showcases the distribution of feature scores for the different features generated by XGBoost. Based on that, we selected a threshold of 0.004 and selected the best features as displayed in Fig 14.

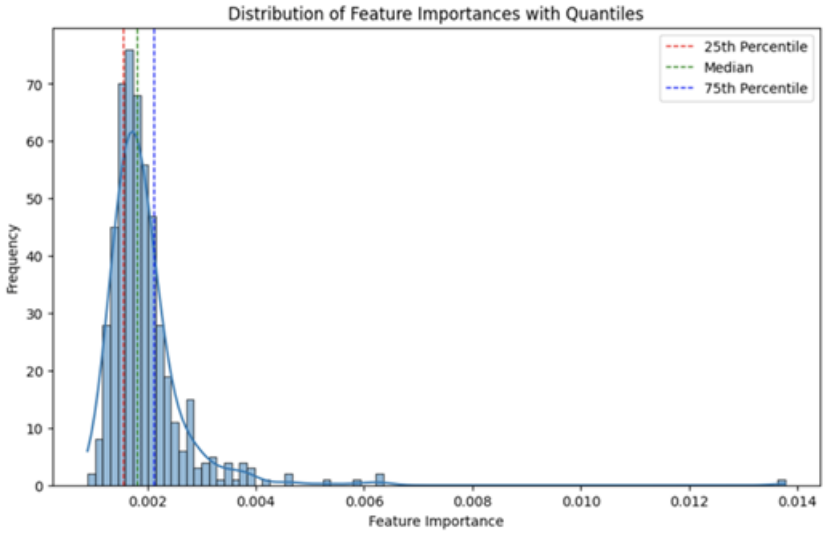

Figure 13: Feature score distribution of XGBoost

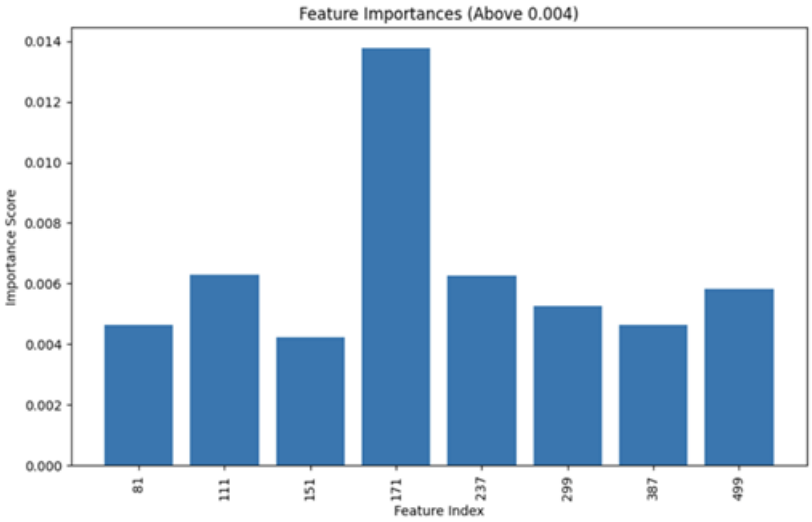

Figure 14: Feature scores of features having > 0.004

Fig. 15 showcases the different scores for the threats, including their precision, recall and f1 scores.

Test Log Loss with best model: 0.517428707046083
Accuracy: 0.8490471414242728

	precision	recall	f1-score	support
0	0.84	0.77	0.80	392
1	0.88	0.61	0.72	244
2	0.87	0.70	0.78	252
3	0.83	0.98	0.90	2006
4	0.92	0.71	0.80	254
5	0.87	0.70	0.77	264
6	0.89	0.77	0.83	388
7	0.94	0.70	0.80	188
accuracy			0.85	3988
macro avg	0.88	0.74	0.80	3988
weighted avg	0.85	0.85	0.84	3988

Figure 15: XGBoost scores for different types of threats

When doing a reverse embedding to generate the highlighted sections in the input threat sentences, we noticed a lot of similarity to the sections highlighted by the LightGBM model as shown in Fig 16.

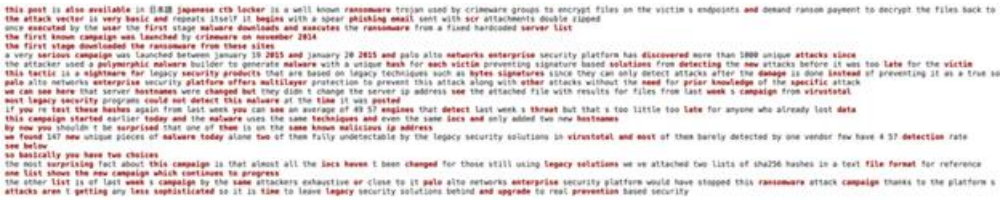

Figure 16: XGBoost model interpretation for different threat descriptions

However, there were some differences. These can be seen in few sentences in Fig 17 wherein XGBoost is able to accurately pinpoint and highlight essential sections / keywords; whereas, LightGBM also captures unnecessary or neighboring words.

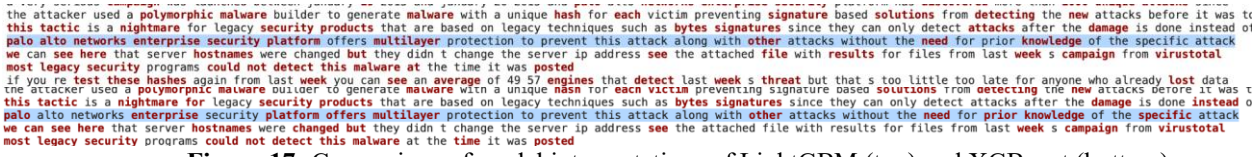

Figure 17: Comparison of model interpretations of LightGBM (top) and XGBoost (bottom)

6. Conclusion

This research presents an integrated, data-driven framework for cyber threat detection, classification, and mitigation, combining the strengths of machine learning, text analysis, and graph-based modeling. Through the use of a multi-modal dataset encompassing network traffic, textual data, and entity relationships, we effectively identified threat patterns, diagnosed threat behaviors, and recommended mitigation strategies. Our application of clustering, relationship analysis, and pattern discovery has revealed meaningful behavioral

similarities among threats and key entities, offering deeper insight into their underlying mechanisms.

The study demonstrates that the fusion of machine learning models with graph-based and textual analysis techniques can significantly enhance the accuracy, automation, and interpretability of cyber threat analysis. This integrated approach not only improves threat detection and diagnosis but also supports the design of more adaptive and proactive cybersecurity strategies.

These findings offer multiple directions for future exploration and refinement of the framework. Future work may focus on the following directions to enhance the robustness, scalability, and adaptability of the proposed framework:

a) **Real-Time Threat Detection and Response**: Extending the framework to support real-time data ingestion and processing will enable live monitoring of network traffic and automated incident response. This involves integrating streaming architectures and deploying low-latency models capable of handling high-throughput environments.
b) **Temporal Modeling of Threat Propagation**: Incorporating temporal graph analytics can allow for the dynamic modeling of threat evolution and propagation over time. This would facilitate the early prediction of attack patterns and enable proactive mitigation strategies based on time-sensitive insights.
c) **Integration of External Threat Intelligence**: Enhancing the system with external threat intelligence feeds—such as dark web data, vulnerability databases, and security bulletins—can improve the detection of emerging and previously unseen threats. This augmentation would also support more comprehensive threat profiling.
d) **Advanced Text Analysis using Deep Learning**: Future iterations of the framework can leverage transformer-based language models (e.g., BERT, RoBERTa, or domain-specific models like SecBERT) to extract richer semantic features from unstructured text, thereby improving the accuracy and contextual understanding in threat classification tasks.
e) **Robustness to Adversarial Attacks and Explainability**: Evaluating and enhancing the resilience of the models against adversarial examples is crucial for security-critical applications. Additionally, incorporating explainable AI (XAI) techniques can help security analysts interpret model decisions, thereby increasing trust and usability in operational settings.
f) **Cross-Domain Generalization and Transfer Learning**: Investigating the applicability of the proposed approach across different sectors (e.g., finance, healthcare, government) and employing transfer learning can help generalize the framework. This would reduce the need for extensive retraining on new datasets and facilitate deployment in varied cybersecurity environments.

Collectively, these directions offer a roadmap for refining the proposed methodology into a fully operational, adaptive, and intelligent cyber defense system capable of addressing the evolving threat landscape.

References

[1] Balisane, H., Egho-Promise, E., Lyada, E., & Aina, F. (2024). Towards improved threat mitigation in digital environments: a comprehensive framework for cybersecurity enhancement. International Journal of Research -Granthaalayah, 12(5). https://doi.org/10.29121/granthaalayah.v12.i5.2024.5655
[2] Camacho, N. (2024). The role of ai in cybersecurity: addressing threats in the digital age. JAIGS, 3(1), 143-154. https://doi.org/10.60087/jaigs.v3i1.75
[3] Devi, A. (2024). Transitioning from reactive to proactive cyber security using machine learning. Int Res J Adv Engg Mgt, 2(08), 2601-2605. https://doi.org/10.47392/irjaem.2024.0378
[4] Fakiha, B. (2023). Enhancing cyber forensics with ai and machine learning: a study on automated threat analysis and classification. International Journal of Safety and Security Engineering, 13(4), 701-707. https://doi.org/10.18280/ijsse.130412
[5] Marinho, R. and Filho, R. (2023). Automated emerging cyber threat identification and profiling based on natural language processing. Ieee Access, 11, 58915-58936. https://doi.org/10.1109/access.2023.3260020
[6] Mohammed, B. (2024). The impact of artificial intelligence on cyberspace security and market dynamics. Brazilian Journal of Technology, 7(4), e74677. https://doi.org/10.38152/bjtv7n4-019
[7] Rangaraju, S. (2023). Ai sentry: reinventing cybersecurity through intelligent threat detection. Eph - International Journal of Science and Engineering, 9(3), 30-35. https://doi.org/10.53555/ephijse.v9i3.211
[8] Rehan, H. (2024). Ai-driven cloud security: the future of safeguarding sensitive data in the digital age. JAIGS, 1(1), 132-151. https://doi.org/10.60087/jaigs.v1i1.89
[9] Sanmorino, A. (2024). The role of data science in enhancing web security. Jeecs (Journal of Electrical Engineering and Computer Sciences), 9(2), 119-116. https://doi.org/10.54732/jeecs.v9i2.4
[10] Sarker, I. (2022). Machine learning for intelligent data analysis and automation in cybersecurity: current and future prospects. Annals of Data Science, 10(6), 1473-1498. https://doi.org/10.1007/s40745-022-00444-2
[11] Shelke, P. and Hämäläinen, T. (2024). Analysing multidimensional strategies for cyber threat detection in security monitoring. European Conference on Cyber Warfare and Security, 23(1), 780-787. https://doi.org/10.34190/eccws.23.1.2123
[12] Abrahams, T., Ewuga, S., Dawodu, S., Adegbite, A., & Hassan, A. (2024). A review of cybersecurity strategies in modern organizations: examining the evolution and effectiveness of cybersecurity measures for data protection. Computer Science & It Research Journal, 5(1), 1-25. https://doi.org/10.51594/csitrj.v5i1.699
[13] Ahsan, M., Nygard, K., Gomes, R., Chowdhury, M., Rifat, N., & Connolly, J. (2022). Cybersecurity threats and their mitigation approaches using machine learning—a review. Journal of Cybersecurity and Privacy, 2(3), 527-555. https://doi.org/10.3390/jcp2030027
[14] Ali, M. and ALQARAGHULI, A. (2023). A survey on the significance of artificial intelligence (ai) in network

cybersecurity. BJN, 2023, 21-29. https://doi.org/10.58496/bjn/2023/004
[15] Cheng, E. and Wang, T. (2022). Institutional strategies for cybersecurity in higher education institutions. Information, 13(4), 192. https://doi.org/10.3390/info13040192
[16] Gülbay, B. and Demirci, M. (2024). A framework for developing strategic cyber threat intelligence from advanced persistent threat analysis reports using graph-based algorithms.. https://doi.org/10.20944/preprints202407.1408.v1
[17] Jamnik, K. (2024). Mathematical frameworks for threat intelligence analysis: leveraging graph theory and machine learning for cyber threat assessment. cana, 31(3s), 503-519. https://doi.org/10.52783/cana.v31.804
[18] Nazir, I. and Tiwari, S. (2023). Impact of machine learning in cybersecurity augmentation., 147-154. https://doi.org/10.48001/978-81-966500-9-4_12
[19] Pandit, R. (2022). A survey on effective machine learning techniques in the field of cyber security. International Journal of Advanced Research in Computer Science, 13(4), 56-61. https://doi.org/10.26483/ijarcs.v13i4.6893
[20] Piplai, A., Mittal, S., Joshi, A., Finin, T., Holt, J., & Zak, R. (2020). Creating cybersecurity knowledge graphs from malware after action reports. Ieee Access, 8, 211691-211703. https://doi.org/10.1109/access.2020.3039234
[21] Tulsyan, R., Shukla, P., Singh, T., & Bhardwaj, A. (2024). Cyber security threat detection using machine learning. Interantional Journal of Scientific Research in Engineering and Management, 08(10), 1-6. https://doi.org/10.55041/ijsrem37949
[22] Wagner, P. and Mapp, W. (2024). Identifying information technology (it) and cybersecurity executives' competencies to support comprehensive cybersecurity programs. European Conference on Cyber Warfare and Security, 23(1), 617-625. https://doi.org/10.34190/eccws.23.1.2317
[23] Cyber Threat Intelligence Dataset: https://www.kaggle.com/code/ramoliyafenil/cyber-threat-classification-using-lstm/input?select=cyber-threat-intelligence_all.csv
[24] Watts, E., Worthington, A., Sheffield, E., & Featherstone, C. (2024). Structural code graphing for automated ransomware detection using novel quantum graph-spectral fingerprinting.. https://doi.org/10.31219/osf.io/cwe3y
[25] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785
[26] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*. https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html
[27] Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (2nd ed.). O'Reilly Media.
[28] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.