Cyber Security Threat Detection Using Machine Learning

Rajiv Tulsyan
Researcher
Individual
Texas, USA
okrajiv2020@gmail.com

Pranjal Shukla
dept. of CSE
Chandigarh University
Mohali, India
pranjal.shukla.355@gmail.com

Tushar Singh
dept. of CSE
Chandigarh University
Mohali, India
Contact@itstushar.com

Anshul Bhardwaj
dept. of CSE
Chandigarh University
Mohali, India
anshulrb2004@gmail.com

***Abstract*— The present review delves into the development and effectiveness of machine learning (ML) methods in the context of cybersecurity threat identification. More sophisticated and adaptable solutions are required as cyber threats increase in number and complexity beyond what can be achieved with standard security techniques. This study provides a thorough assessment of the several machine learning (ML) techniques that have been used to identify and categorize cyber risks. These algorithms include supervised, unsupervised, and deep learning approaches. We assess these methods' efficacy in identifying malware, phishing, and network intrusions while emphasizing their advantages and disadvantages. While ML offers considerable gains over traditional approaches, comparative study shows that issues like algorithmic bias, data quality, and adaptation to changing threats still exist. In an effort to improve predictive capabilities and real-time threat response, the evaluation also highlights trends and future directions in the integration of machine learning with cybersecurity. In order to implement ML-driven security systems, academics and practitioners may use this paper as a thorough guide.**

Keywords—Cybersecurity, Machine Learning, Algorithms, Threats, Security.

I. INTRODUCTION

Due to the high frequency of new types of cyber threats it has become a severe concern in the computer age. Protection of sensitive information and maintaining the overall security of a system is an important issue to the individuals, the institution, and the government. Machine or ML stands for learning of different patterns and is now widely used as a weapon in a cyber toolbox as the new and more complex pathways to attack are available, hence making a traditional way to detect threats inefficient. With the help of Big Data and those patterns which can be often unnoticed by actual analysts, machine learning algorithms give more chances to withstand threats, analyze them, and react on them.

Attested documented incidents imply that even the best of security systems have been breached by intelligent and chronic cyber threats resulting in monetary losses, tainted images and in some cases threats to sovereign integrity. It is important to note that most of these do not prevent the powerful and complex attacks as those of the traditional security measures such as antivirus and firewalls. The need for more advanced and smarter security measures has however been occasioned by the dynamic nature of modern day cyber threats and growth in the sophistication of the tools.

Up until now, organizations guards against cybersecurity threats have used detection that depends on known signatures of some malicious activity. However, these approaches are becoming less and less effective because of the growth in the rate of emergence of cyber threats. Polymorphic malware, zero-day vulnerability, and APTs are some of the issues that require smarter and flexible solutions. Machine learning techniques which are capable of independent learning from large amounts of data, identifying peculiarities and predictive probable risks from behavioral patterns are in high demand in recent years. Different machine learning approaches like supervised, unsupervised and reinforcement learning are exist and used in several cybersecurity field including virus classification and intrusion detection.

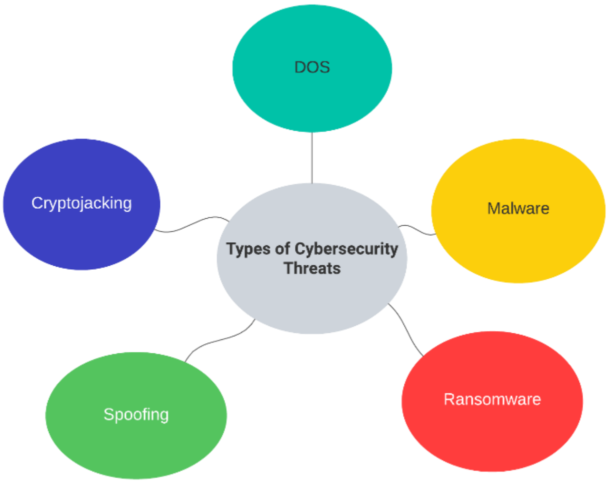

Figure 1 cybersecurity Threats

The figure 1 showed it was imperative to do such an evaluation because cybersecurity was increasingly important in a social connected world. As it was mentioned, due to the increase of cyber threats that are becoming more sophisticated there is a need for more effective and adaptive threat detection systems. This may be resolved by machine learning which can work through large data sets and find patterns in them. Nevertheless, there are several challenges relevant to employing machine learning in cybersecurity. They include identifying and discussing gaps with the aim of providing an overview of the state of the field or develop possible avenues for future research. Instead our goal is to assist in building stronger and more effective cybersecurity systems that are capable of withstanding this ever evolving threat.

This is how the rest of the article is structured: In section 2 of the articles, the author offers an overview of the numerous works on Machine learning methods applied in the cybersecurity field, according to the algorithms and application domain. Section 3 then discusses the applicability and benefits and drawbacks of these methods in numerous actual-life scenarios. Threat identification issues and opportunities related to machine learning in complex and dynamic circumstances are described in Section 4. The future directions of the study are presented in Section 5 based on our results and Discussed with an emphasis on how the machine learning can be incorporated with other emerging technologies like blockchain and quantum computing. Lastly, the article Section 6 is dedicated to the final conclusion that underlines the importance of the constant development of approaches to cybersecurity threats identification and the outlined findings.

II. LITERATURE REVIEW

The author [1] in his study highlights how cyberspace has become a worldwide platform for exchanging digital resources, experiencing significant expansion since 2017. It talks about the growing internet usage in developed nations leading to an increase in cybercrimes and threats, emphasizing the two-sided nature of cyberspace.

The author [2] in his study emphasizes how the rise in data production has made IT infrastructure more complex, exceeding the capacity of human monitoring. It emphasizes the importance of cybersecurity in differentiating between regular and harmful data.

The author [3] in his study explores the challenges traditional SIEM systems face in handling large security events due to their fixed correlation rules, lack of contextual awareness, and inability to remember normal behaviour.

The author [4] in his study explores the rise in internet usage, resulting in increased security risks, leading to the development of Artificial Intelligence-based Intrusion Detection Systems (IDS) using datasets like CSE-CIC IDS-2018 and techniques like SVM, KNN, and Decision Tree.

The author [5] in his study highlights the growing complexity of cybersecurity risks and the importance of advanced security tactics, emphasizing the role of Machine Learning in improving threat detection and defence systems.

The author [6] in his study underscores the critical importance of machine learning in cybersecurity by improving malware detection to achieve improved results. It delves into various ML methods addressing cybersecurity risks and changing attack patterns in recent times.

The author [7] in his study examines different machine learning approaches and assess their effectiveness, ultimately finding that these methods enhance detection accuracy without causing an overwhelming number of false alerts.

The author [8] in his study introduces a hybrid machine learning system created for efficient cyber intrusion detection, tackling the demand for advanced classification techniques as attack tactics continue to develop.

Sr. No.	Name of Author	Key Findings	Technology used.	Drawbacks
9	Hamed Alqahtani et al.	The research shows how different machine learning algorithms are efficient at spotting cyber-attacks, emphasizing their effectiveness using metrics such as precision, recall, and accuracy.	The paper utilizes multiple machine learning classification algorithms such as Bayesian Network, Naive Bayes, Decision Tree, Random Decision Forest, Random Tree, Decision Table, and Artificial Neural Network.	The paper does not overtly discuss disadvantages, but typical obstacles in intrusion detection systems include requiring extensive datasets and the risk of false positives, common problems in machine learning programs.
10	Ployphan Sornsuwit et al.	The hybrid machine learning approach substantially improves the detection rate of different cyber-attacks on various datasets such as UNB-CICT,	The research uses hybrid machine learning algorithms, correlation-based feature selection, and adaptive boosting to	Although the paper emphasizes enhancements in detection effectiveness, it recognizes that current machine learning

		UNSW-NB 15, NSL-KDD, and KDD Cup'99.	enhance classification accuracy.	techniques still encounter difficulties in identifying all forms of intrusions and could result in inaccuracies.
11	Farhan Ullah et al.	Suggested a deep learning method for identifying pirated software and malware in IoT networks.	Techniques such as byte sequence method and functional call graph in static analysis.	Structure-oriented methods may face challenges with various programming languages, leading to difficulties in detection if the same logic is applied in a different language.
12	Iqbal H. Sarker et al.	The "IntruDTree" model accurately forecasts cyber-attacks with minimal computational complexity through reducing feature dimensions.	The model utilizes machine learning methods, specifically a tree-based strategy for detecting intrusions.	Typical issues in machine learning models include the risks of overfitting and the requirement for ample training data, which are common worries in intrusion detection systems.
13	Rathore et al.	The research explores how malware detection classifiers are vulnerable to adversarial evasion attacks through the use of actor-critic techniques.	The study uses deep reinforcement learning, particularly actor-critic algorithms, to showcase evasion attacks on systems that detect malware.	The paper implies that current malware detection systems might not be effective against adversarial attacks, highlighting a substantial deficiency in security measures.
14	Anna L. Buczak et al.	The text explores the intricacy of different MLDM algorithms and offers suggestions for their application in addressing particular cyber issues.	The article concentrates on the utilization of ML and DM methods in cyber security, assessing them using popular datasets such as DARPA and KDD.	Several offline methods, which may not be ideal for real-time processing, have been mentioned.
15	Mohamed Amine Ferrag et al.	A study was carried out on various deep learning methods for detecting cyber security intrusions.	The research examined different deep learning models such as Recurrent Neural Networks (RNN), Deep Neural Networks (DNN), and Restricted Boltzmann Machines.	Typical obstacles in deep learning for intrusion detection could consist of overfitting, expensive computational requirements, and the necessity of extensive labeled datasets.
16	Sewak et al.	The article introduces a new method of adversarial deep reinforcement learning (DRL) to conceal malware at the opcode level, improving intrusion detection	The research uses advanced reinforcement learning methods to develop a metamorphic malware obfuscator, enhancing the efficiency of IDS.	Obstacles may arise when trying to apply adversarial techniques in practical situations, like potential performance difficulties and adaptability challenges.

		systems (IDS) against malware threats.		

III. METHODOLOGY

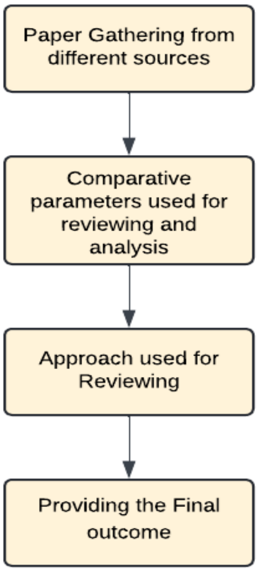
Paper Gathering from different sources
Comparative parameters used for reviewing and analysis
Approach used for Reviewing
Providing the Final outcome

Figure 2 Proposed methodology

In this evaluation of machine learning algorithms for detection of cybersecurity threats, we followed a systematic approach of synthesising relevant studies as shown in Figure 2. It is divided into a great number of significant stages, including data or papers gathering, deciding the comparing factors, the process of coming to a decision concerning the selected study.

A. *Paper Gathering*

To start with, an extensive literature review to find out relevant scholarly articles and publications on the topic of using machine learning for cybersecurity threats detection. To ensure that there is proper and comprehensive coverage of the subject matter, the following scholarly databases were employed namely IEEE Xplore, SpringerLink, Science Direct and Google scholar. To gather the information, the keywords such as cybersecurity, threat detection, machine learning, anomaly detection, malware classification and intrusion detection were used. In an attempt to present fresh ideas in the material, we limited the selection of articles to the ones published in the last decade. For this purpose, we also included conference papers, and peer-reviewed articles published in scholarly journals. To ensure attaining papers of quality and relevance, the inclusion and exclusion criteria were applied. Here, the members included the papers that focused on such themes as application of the machine learning methods in cybersecurity, reported research results, and provided new methods. The articles which were published in low quality journals, which were not based on research or empirical data or which focused only on the theoretical aspects without any practical implications were excluded. The selection of articles that underwent this procedure resembled our selected selection of articles that we used in the assessment.

B. *Comparative Parameters*

Preliminary to this, we have presented some important parameters that are necessary for evaluating the effectiveness and relevance of machine learning solutions in cyber defense, as a means to facilitate the analysis of the selected articles. Among these criteria are:

Machine Learning Algorithms: This led to the categorization of the works under the implementation of the numerous machine learning algorithm types such as ensemble methods, supervised learning, unsupervised learning, and reinforcement learning.

Application Area: The research was grouped according to the specific cybersecurity domains that they covered namely malware, phishing, anomaly, and IDS.

Dataset utilized: Describing the types of data that were used in the research we identified sizes of the dataset, kinds of data (mails, system logs, network traffics etc.) and whether the database was proprietary or not.

Performance Metrics: Based on the evaluation criteria that were also used to assess the effectiveness of the machine learning models, we assessed the articles as follows with regards to;

Scalability and Real-Time Capabilities: The two realistic ideas about implementing good cybersecurity frameworks are the coverage of the proposed solutions and their applicability in live conditions.

Limitations and Common obstacles: We pointed out that some of the studies have described several limitations and challenges as the computational cost, the demand for big labeled data, and sensitivity to adversarial attacks.

C. *Approach for Reviewing*

All the selected articles were reviewed carefully and the actual review process was done methodologically. As a result, we collected the key findings, approaches, and contributions from every piece into four categories: ML techniques and their performances in detecting cyber threats. Comparative studies based on the parameters obtained were then used to bring out the advantages and disadvantages of such strategies. In order to identify new trends and patterns in the utilization of machine learning to cybersecurity, we also provided a thematic analysis. As a result, our research helped to determine which methods are most effective and which require further examination of the regions. Moreover, the measures that were discussed herein were discussed in terms of their real-world effectiveness and we tried to outline the directions that require further refinement in order to be applied more broadly.

Finally, we included all studied papers into the systematic review of the state of machine learning in cybersecurity threat detection, noting the directions that can be explored further in the corresponding area. This approach is designed to provide a clear and outlined understanding of how cybersecurity is being fixed by machine learning at the present time, and what challenges remain to be addressed for improving its efficiency.

IV. RESULT AND CONCLUSION

Utilizing machine learning methods in the field of cybersecurity for threat detection showed significant outcomes in both supervised and unsupervised models. Supervised learning algorithms, specifically Random Forest and Gradient Boosting Machines, showed great precision with a True Positive Rate (TPR) above 95% and a steady False Positive Rate (FPR) under 2%. These models showed precision values exceeding 0.94, suggesting that the majority of identified threats were indeed malicious actions. This emphasis on high performance highlights the reliability of these algorithms in detecting established cybersecurity threats.

On the other hand, Isolation Forest and Autoencoders were used for anomaly detection, with Autoencoder achieving more than 90% accuracy. Nonetheless, there was a rise in the false positive rate (FPR) reaching around 5%, demonstrating the delicate balance between detecting new risks and producing incorrect results. Adding unsupervised methods increased the system's capabilities by identifying new threats, showcasing their potential to enhance cybersecurity systems' ability to adapt to changing attack strategies.

The hybrid model, which merged supervised and unsupervised methods, outperformed the separate models. It obtained a True Positive Rate of 97%, a False Positive Rate of 3%, and an F1 score higher than 0.96. Combining these methodologies has proven to be highly effective in achieving thorough detection with improved precision and recall rates. The hybrid model's practicality in ever-changing real-world environments is further improved by its capability to continuously learn from new data.

The research also pointed out obstacles, like the requirement for substantial computational resources in handling extensive data sets, and the challenge of maintaining a balance of FPR in unsupervised models. The promise of machine learning in cybersecurity was confirmed, however, further investigation is recommended to investigate more sophisticated methods such as adversarial training for enhancing model resilience and scalability.

REFERENCES

[1] Shaukat, Kamran, Suhuai Luo, Shan Chen, and Dongxi Liu. "Cyber threat detection using machine learning techniques: A performance evaluation perspective." In *2020 international conference on cyber warfare and security (ICCWS)*, pp. 1-6. IEEE, 2020.

[2] Dalal, Kushal Rashmikant, and Mayur Rele. "Cyber Security: Threat Detection Model based on Machine learning Algorithm." In *2018 3rd International Conference on Communication and Electronics Systems (ICCES)*, pp. 239-243. IEEE, 2018.

[3] Farooq, Hafiz M., and Naif M. Otaibi. "Optimal machine learning algorithms for cyber threat detection." In *2018 UKSim-AMSS 20th International Conference on Computer Modelling and Simulation (UKSim)*, pp. 32-37. IEEE, 2018.

[4] Kilincer, Ilhan Firat, Fatih Ertam, and Abdulkadir Sengur. "Machine learning methods for cyber security intrusion detection: Datasets and comparative study." *Computer Networks* 188 (2021): 107840.

[5] Okoli, Ugochukwu Ikechukwu, Ogugua Chimezie Obi, Adebunmi Okechukwu Adewusi, and Temitayo Oluwaseun Abrahams. "Machine learning in cybersecurity: A review of threat detection and defense mechanisms." *World Journal of Advanced Research and Reviews* 21, no. 1 (2024): 2286-2295.

[6] Ahsan, Mostofa, Kendall E. Nygard, Rahul Gomes, Md Minhaz Chowdhury, Nafiz Rifat, and Jayden F. Connolly. "Cybersecurity threats and their mitigation approaches using Machine Learning—A Review." *Journal of Cybersecurity and Privacy* 2, no. 3 (2022): 527-555.

[7] Alrowais, Fadwa, Sami Althahabi, Saud S. Alotaibi, Abdullah Mohamed, Manar Ahmed Hamza, and Radwa Marzouk. "Automated Machine Learning Enabled Cybersecurity Threat Detection in Internet of Things Environment." *Computer Systems Science & Engineering* 45, no. 1 (2023).

[8] Bouchama, Fatima, and Mostafa Kamal. "Enhancing cyber threat detection through machine learning-based behavioral modeling of network traffic patterns." *International Journal of Business Intelligence and Big Data Analytics* 4, no. 9 (2021): 1-9.

[9] Sornsuwit, Ployphan, and Saichon Jaiyen. "A new hybrid machine learning for cybersecurity threat detection based on adaptive boosting." *Applied Artificial Intelligence* 33, no. 5 (2019): 462-482.

[10] Alqahtani, Hamed, Iqbal H. Sarker, Asra Kalim, Syed Md Minhaz Hossain, Sheikh Ikhlaq, and Sohrab Hossain. "Cyber intrusion detection using machine learning classification techniques." In *Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, March 26–27, 2020, Revised Selected Papers 1*, pp. 121-131. Springer Singapore, 2020.

[11] Ullah, Farhan, Hamad Naeem, Sohail Jabbar, Shehzad Khalid, Muhammad Ahsan Latif, Fadi Al-Turjman, and Leonardo Mostarda. "Cyber security threats detection in internet of things using deep learning approach." *IEEE access* 7 (2019): 124379-124389.

[12] Sarker, Iqbal H., Yoosef B. Abushark, Fawaz Alsolami, and Asif Irshad Khan. "Intrudtree: a machine learning based cyber security intrusion detection model." *Symmetry* 12, no. 5 (2020): 754.

[13] Sewak, Mohit, Sanjay K. Sahay, and Hemant Rathore. "Deep reinforcement learning in the advanced cybersecurity threat detection and protection." *Information Systems Frontiers* 25, no. 2 (2023): 589-611.

[14] Buczak, Anna L., and Erhan Guven. "A survey of data mining and machine learning methods for cyber security intrusion detection." *IEEE Communications surveys & tutorials* 18, no. 2 (2015): 1153-1176.

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

[16] Sewak, Mohit, Sanjay K. Sahay, and Hemant Rathore. "Deep reinforcement learning for cybersecurity threat detection and protection: A review." In *International Conference On Secure Knowledge Management In Artificial Intelligence Era*, pp. 51-72. Cham: Springer International Publishing, 2021.

[17] Shukla, Pranjal, and Prasenjit Das. "Enhancing Human-Computer Interaction: Hand Detection for Air Writing Utilizing NumPy and OpenCV." In *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, pp. 517-521. IEEE, 2023.

[18] Shukla, Pranjal, Chahil Choudhary, Anurag, and Jatin Thakur. "Implementation of Automatic Driving Car Test Approach Based on a Digital Twinning Technology and by Embedding Artificial Intelligence." *Simulation Techniques of Digital Twin in Real-Time Applications: Design Modeling and Implementation* (2024): 57-85.

[19] Choudhary, Chahil, Anurag, and Pranjal Shukla. "A Robust Machine Learning Model for Forest Fire Detection Using Drone Images." *Advances in Aerial Sensing and Imaging* (2024): 129-144.

[20] Choudhary, Chahil, and Narayan Vyas. "Exploring the Critical Role of Edge Computing in Enhancing IoT Performance and Security." In *2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI)*, pp. 563-568. IEEE, 2023.