Reza Rizvi
Professor
Department of Mechanical Engineering
York University

Dear Professor Rizvi,

I have prepared the following report “Machine Learning for Cybersecurity” detailing a possible solution for the Engineering Grand Challenge of Securing Cyberspace. This report will be ready for publishing by March 22, 2022.

In this report I discuss the Engineering Grand Challenge of Securing Cyberspace and its associated problems. I then offer as a solution the development of machine learning for cybersecurity, detailing its capabilities, strengths, weaknesses, and effectiveness in making progress for the grand challenge. I intend for this report to educate and inform you of the importance of these issues and the suitability of the new technology.

I encourage you to contact me if you have questions, concerns, or any feedback. Thank you for your time.

Sincerely,

FirstName LastName

# Machine Learning for CyberSecurity

## A Solution to the Engineering grand Challenge of Securing Cyberspace

March 8, 2022

I attest that that this is my original work.

# Executive Summary

# Contents

# 1 Introduction and Background

Technology and the internet have been growing more and more pervasive over the last few decades, and this shows no sign of stopping. This comes with many benefits, like an abundance of readily available information, almost instantaneous communication, and the entire industry of online markets. However, this also comes with consequences. Perhaps one of the most pressing issues with the growth of technology is the introduction of a new type of malicious activity — cybercrime.

Cybercrime is the use of the Internet for illegal purposes. This includes crimes targeting individuals, such as malware, ransomware, phishing, and identity theft. But with the increasing dependance of industry on the internet, this cybercrime includes crimes that target populations and large systems. For example, cyberattacks can target power grids, military systems.

The engineering grand challenge of securing cyberspace seeks to address this problem. The internet is one of the most complex systems ever engineered.

Such cyberattacks have happened. On June 1, 2020, the University of California, San Francisco, was hacked by a ransomware campaign that threatened to release confidential information, to which the university paid approximately $1.14 million to the group [1]. In fact, in 2015, cybercrime was estimated to have had a global cost of $3 trillion [2], and is expected predicted to reach $10.5 trillion by 2025 [3].

As is evident by successful cyberattacks of reputable organizations, current methods of dealing with these attacks are insufficient. The main method of developing cybersecurity systems is finding fixes for vulnerabilities only when they are unfortunately discovered.

# 2 Machine Learning's Use in Cybersecurity

Traditionally, cybersecurity algorithms were written manually from heuristics [4]. But the rapid growth of the internet and technology in general has led to constantly changing cybersecurity threats. As a result, these manually written heuristic algorithms are insufficient — they cannot keep up with the evolving threats [4]. Machine learning offers a solution to this problem.

Machine learning models are able to "learn" certain data patterns to predict behavior [4]. In this case, their goal is to predict whether some online activity is malicious or legitimate. To accomplish this, the model must be trained with training data and tested to ensure it is effective. This first involves data-driven tasks, like gathering and cleaning data [4]. This data can then be used to train the model, which may take seconds to days, depending on the algorithm chosen [5]. Once the model is trained, it must be tested to ensure it is accurately detecting malicious activity, which also takes a variable amount of time depending on the machine learning algorithm chosen [5]. Figure 1 shows a diagram of this process.

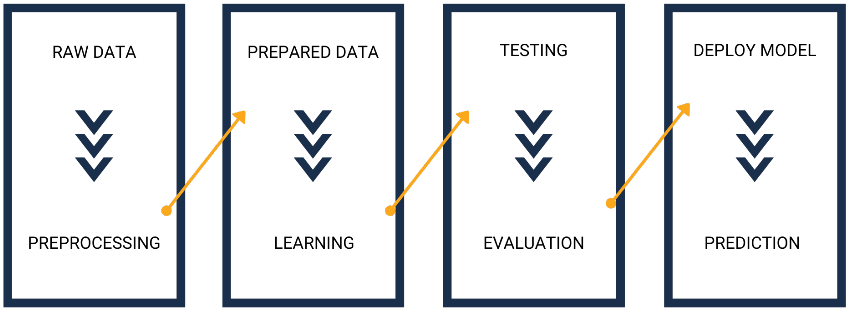

**Figure 1:** A Diagram of a Typical Machine Learning Process. [6]

# 3 Evidence of Machine Learning's Effectiveness

Machine learning algorithms are already being employed to detect cyberattacks. The following are a few case studies of machine learning being successfully implemented.

## 3.1 MalDozer—Automatic Malware Detection for Android

Android, an open-source mobile device operating system, is particularly vulnerable to malware, in part due to its openness [7]. In response, multiple machine learning technologies are being developed to protect Android devices from malware [7]. One such technology is called MalDozer. Several experiments were conducted by presenting MalDozer with datasets containing varying amounts of malware [7]. The experimenters found that its accuracy rate was 96%-99% with a false positive rate of 0.06%-2% [7].

## 3.2 Windows Defender Antivirus

The Windows Defender Antivirus is not just being tested, but is actually put in use and has intervened in cyberattaaks. In 2018, a new malware attack campaign was launched against over a thousand users of Windows 7 Pro [8]. The Windows Defender Antivirus features lightweight machine learning models built into the client, which responded immediately to the attack [8]. These models detected a high probability of maliciousness, so they sent data to the Windows Defender Antivirus cloud protection service, which runs more complex machine learning models [8]. Through this, the cloud protection service correctly identified the requests as a cyberattack and responded back to the clients, instructing them to block the attack [8]. The use of machine learning algorithms were able to protect thousands of users from a cyberattack with no human intervention.

# 4 Drawbacks of Machine Learning for Cybersecurity

In its current state, machine learning has many drawbacks for use in cybersecurity, some of which make it infeasible for organizations to use.

## 4.1 Availability of Datasets

Since cyberattacks can be complex and varied, large datasets are needed to train the machine learning models to ensure they can protect against all attacks. This is not the case with the existing datasets available. Current datasets contain lots of old data and redundant information [5]. This can be somewhat improved after cleaning the data, but even then there is the issue of volume — there is not enough data to properly train the models [5]. As a result, the machine learning models are not totally equipped for identifying new cyberattacks. This also introduces a barrier of entry, as larger organizations may be able to work around these issues, but smaller organizations do not have the resources to do so.

## 4.2 Lack of Research and Adoption

While the field of machine learning receives lots of research, this research is mainly focused on deep-learning algorithms for applications like self-driving cars [9]. Machine learning for cybersecurity purposes has yet to receive this same amount of attention. Due to this lack of research, widely adopted machine learning models for cybersecurity are limited, using mostly rule-based techniques [9]. Furthermore, this lack of research introduces inconsistency across organizations [9]. To be most effective, cybersecurity models need to have consistent behavior for any attack that may occur. This requires cooperation and research to keep all parts of the internet secure.

# 5 Discussion

# 6 Conclusion

# References

[1] D. Winder, “The University Of California Pays $1 Million Ransom Following Cyber Attack,” *forbes.com*, Jun. 29, 2020. [Online]. Available: https://www.forbes.com/sites/daveywinder/2020/06/29/the-university-of-california-pays-1-million-ransom-following-cyber-attack/?sh=5628ae8618a8 [Accessed March 6, 2022].

[2] Microsoft Secure Blog Staff, “The Emerging Era of Cyber Defense and Cybercrime,” *Microsoft*, Jan. 27, 2016. [Online]. Available: https://www.microsoft.com/security/blog/2016/01/27/the-emerging-era-of-cyber-defense-and-cybercrime/ [Accessed March 7, 2022].

[3] S. Morgan, “Cybercrime To Cost The World $10.5 Trillion Annually By 2025,” *cybersecurityventures.com*, Nov. 13, 2020. [Online]. Available: https://cybersecurityventures.com/cybercrime-damages-6-trillion-by-2021/ [Accessed March 7, 2022].

[4] I.H. Sarker, A.S.M. Kayes, S. Badsha, “Cybersecurity data science: an overview from machine learning perspective,” J Big Data, Jul. 1, 2020. [Online]. Available: https://doi.org/10.1186/s40537-020-00318-5 [Accessed February 28, 2022].

[5] Y. Xin et al., “Machine Learning and Deep Learning Methods for Cybersecurity,” *IEEE Access*, vol. 6, pp.3 5365-35381, 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8359287/citations?tabFilter=papers [Accessed March 7, 2022].

[6] Echosec Systems, “How Is Machine Learning Used in Cybersecurity?,” *Echosec Systems*. [Online]. Available: https://www.echosec.net/blog/how-is-machine-learning-used-in-cybersecurity [Accessed March 7, 2022].

[7] Dongliang Chen, Paweł Wawrzynski, Zhihan Lv, “Cyber security in smart cities: A review of deep learning-based applications and case studies,” *Sustainable Cities and Society*, Volume 66, Mar. 2021. [Online]. Available: https://doi.org/10.1016/j.scs.2020.102655 [Accessed March 8, 2022].

[8] Microsoft Defender Security Research Team, “How artificial intelligence stopped an Emotet outbreak,” *Microsoft*, Feb. 14, 2018. [Online]. Available: https://www.microsoft.com/security/blog/2018/02/14/how-artificial-intelligence-stopped-an-emotet-outbreak/ [Accessed March 7, 2022].

[9] K. Bresniker, A. Gavrilovska, J. Holt, D. Milojicic, T. Tran, “Grand Challenge: Applying Artificial Intelligence and Machine Learning to Cybersecurity,” *Computer*, vol. 52, no. 12, pp. 45-52, Dec. 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8909930 [Accessed March 8, 2022].