Machine Learning/AI in Cybersecurity: Literature Review

Christopher Wong

Abstract

Introduction

Cybersecurity is a field within Computer Science that focuses on preventing digital attacks and threats. While it most commonly encompasses security on digital software, physical security is also an important component, such as keeping important hardware safe. Cybersecurity in the digital realm begins at design - most network devices are designed to be naturally secure for their users. However, specialized cybersecurity personnel are required for maintenance and defense against modern-day threats. Cybersecurity threats are defined as any unauthorized breach into a computer system or network. While vandalism and thrill-seeking can be attributed to some attacks, many attacks aim to steal sensitive data, or maliciously shut down infrastructure, usually for financial gain. Thus the importance of the cybersecurity field can be quickly recognized; the more entrenched digital technology becomes in our everyday lives, the more cybersecurity must be taken seriously. According to several of the articles I have parsed as a part of my literature review, cybersecurity threats have increased exponentially in the past decade, and will likely continue to do so in coming years.

As mentioned, cybersecurity begins at the base-design of network devices and computers. However, the threats and dangers of modern cyberspace demand specialized cybersecurity knowledge to properly maintain networks, and actively defend against threats. Cybersecurity as a profession has become one of the most prolific subfields of the Computer Science field; due to that the entirety of the industry relies on privacy, and security for the client, just as would any other industry. Historically, cybersecurity relies on basic prevention systems such as Firewalls Or Intrusion Detection Systems. These systems rely on signature-based detection which identify patterns/signatures from a database of known threats. As noted through my research, they are ineffective at detecting zero-day threats, since they are unable to recognize a previously unknown threat. Zero-day viruses and worms can cause irreversible damage that cannot be detected by standard Intrusion Detection Systems. Thus, the ability to detect zero-day threats cannot be understated.

Artificial Intelligence is the broad term for simulated cognition. A more specific term that relates to my topic is Machine Learning. Machine Learning is defined as a computer skill that adapts to information and learns with a live dataset(constantly adapting to user input as part of the dataset); the machine learned program would then be able to make accurate decisions based off The market for Machine Learning tools has exploded in the past half-decade as its viability to be cross-implemented into other industries has been realized.

One of my primary research studies that I reviewed tested the viability of a Machine Learning Model in specifically detecting DDOS(Distributed Denial of Service) Attacks. It provided an in depth look into the training of a neural network dataset and a practical trial of the neural network's performance in a network traffic environment similar to real-world conditions. I plan to do something similar, but I want to extend the dataset to include other detection properties that detect the propagation of malicious

software; so this particular source (Tymoshchuk[1]), proved to be valuable literature to the research that I aimed to conduct.

Materials and Methods

My literature review will focus on four primary studies, plus a few magazine articles that I used for informal contextualization. My primary tools for search were Google Scholar and JSTOR - a resource which I had learned about earlier in the school year. I searched through these databases using the terms, “Cybersecurity”, “Machine Learning” etc. My search covered publications through June 2025. Inclusion criteria was generally wide, but only focused on English language articles. I initially had trouble finding sources that focused on supervised methods, so I had to widen my search.

Results

I explored a total of 4 main studies, as well as 2 magazines which I utilized for an informal contextualization of a few important concepts and developments. These articles covered cybersecurity, AND/OR Machine Learning and AI. Some of the sources covered exclusively either topic, while some focused on the bridge between the two. In my later research, I focused on neural networks and explored their possibility as an effective ML option to classify network traffic in real time.

According to most sources, the number of cybersecurity threats have skyrocketed in the past decade. As digital technology becomes more entrenched in every-day life, digital cyberspace becomes more appealing as a lucrative environment for data theft, and other threats. In particular, a computer science magazine[2] cites growing sophistication in cyberattack technology, and growing state-sponsored cyberterrorism on the international scene. Thus, it can be asserted the necessity of ML/Artificial Intelligence as a necessity to combat the continuing growth of cybersecurity threats. In order to make a proper comparison in a practical environment, traditional IDS(Intrusion Detection Systems) has to be explored. IDS’s inadequacy in a modern context can be seen in this passage:

> IDSs are usually hybrid and have anomaly detection and misuse-detection modules. The anomaly detection module classifies attack patterns with known signatures or extracts new signatures from the attack-labeled data comping from the anomaly module. Often, an anomaly detector is based on a clustering method. Among clustering algorithms, density-based methods … are the most versatile, easy to implement, less parameter or distribution dependent, and have high processing speeds. In anomaly detectors, one-class SVMs also perform well … . Among misuse detectors, because the signatures need to be captured, it is important that the classifier be able to generate readable signatures, such as branch features in a decision tree, genes in a genetic algorithm, rules in Association Rule Mining, or sequences in Sequence Mining. Therefore, black-box classifiers like ANNs and SVMs are not well suited for misuse detection. Several state-of-the-art ML and DM algorithms are suitable for misuse detection. Some of these methods are statistical such as Bayesian networks and HMMs; some are entropy-based such as decision trees; some are evolutionary such as genetic

---

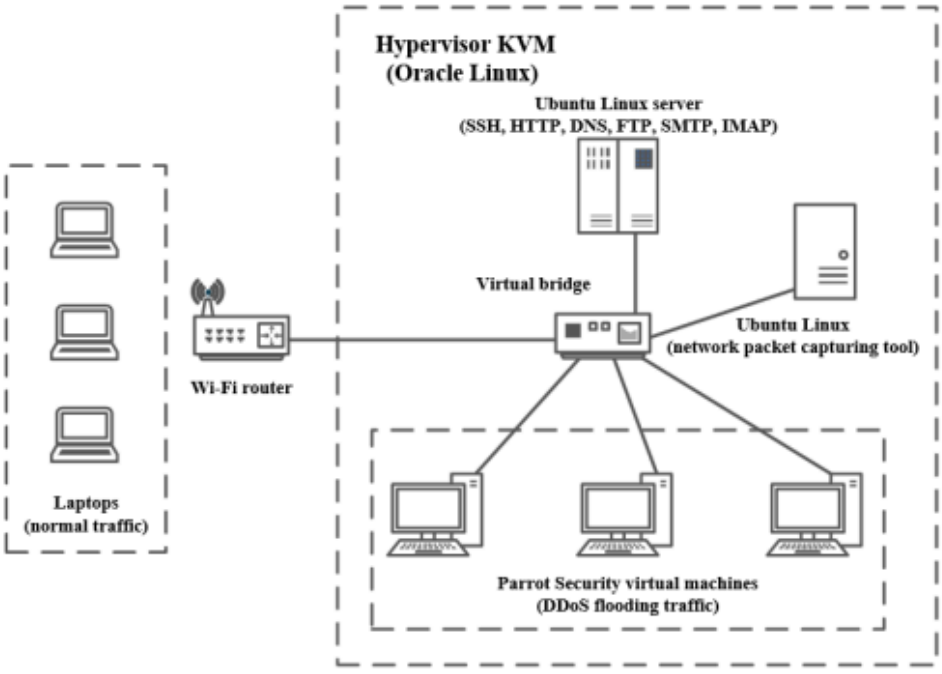
[1] Bellato E, Marini E, Castoldi F, et al. Fibromyalgia syndrome: etiology, pathogenesis, diagnosis, and treatment. Pain Res Treat. 2012;2012:426130. doi: 10.1155/2012/426130. Epub 2012 Nov 4. Erratum in: Pain Res Treat. 2013;2013:960270. [DOI] [PMC free article] [PubMed] [Google Scholar]

[2] Cybermagazine bruh

> algorithms; some are ensemble methods like Random Forests; and some are based on association rules. … methods like Bayesian networks or HMMs may not be the strongest approach because the data do not have the properties that are the most appropriate for them. Evolutionary computation methods may take a long time to run and therefore may not be suitable for systems that train online. If the training data are scarce, Random Forests might have an advantage. If the attack capture is important, decision trees, evolutionary computation, and association rules can be useful. (p. 1173)

This review of traditional IDS systems asserts that even though IDS are generally simple and easy to implement, AI/Machine Learning represent a much more efficient solution to cybersecurity threats, especially in an increasingly hostile digital environment where signature-based and anomaly-based detection can be ineffective for a number of reasons as mentioned.

How exactly a Machine-Learning cybersecurity defense could be implemented remains vague. The implementation described by Tymoshchuk in his study[3] of a Machine Learning model to detect DDOS attacks on a network included a virtual bridge on a WIFI router within a simulated VM (virtual machine). The provided network is:s



3

The process described by this particular study proposed a four-step approach taking place on the virtual bridge.

1. Generation of normal & malicious traffic
2. Capturing networking packets
3. Feature extraction from the captured packets
4. Classification of a neural network

Cornell Uni's DDOS model utilizes a neural network as part of the Machine-Learning process, complete with 24-106-5 feed-forward model, providing a 24 neuron input layer, a 106 neuron hidden layer, and a 5 neuron output layer. Because of their broad amount of data that they had collected, and the depth and complexity of the neural network, their model produced 99% accuracy in practical trials. This implementation of a Machine Learning model in a practical cybersecurity environment proved to be successful.

From this particular study, it is asserted that neural networks provide the most accurate, and time-effective option for identifying threats. Neural networks are built to simulate the structure of a set of human neurons and make accurate judgements. The neuron within each network receives a basic input, processes it, and transmits the results to the output. There are three main features in a neural network, containing input neurons, hidden neurons, and output neurons. In the most common arrangement, forward-propagation, the processed information moves forward through the three stages. The basic mathematical formula for a single digital neuron is represented as:

$$y = \varphi\left(\sum_{i=1}^{n} \omega_i \cdot x_i + b\right)$$

Where:

Xi are the input values, wi are the weight values associated with each input, b is the activation function, n is the number of input signals, and y is the output of that particular neuron. I represents each singular data input that might be a part of the data processed by the neuron as a whole. Wi, or weight can be learned by the machine to represent a different weight or importance

Another topic of importance which is essential to cybersecurity network traffic are supervised methods. Supervised methods include algorithms in which the computer can automatically mark traffic, for example, "malicious", or "benign". These INCLUDE neural networks, the model used by Cornell in the study discussed above. Techniques to test the accuracy of these models over the same dataset in the same environment are called benchmark studies. These metrics will focus on accuracy, precision, and recall over the separate supervised methods. Benchmarking in these situations will help recognize which supervised method is the most effective at a given detection task.

Common supervised methods:

Method	Description	Pros	Cons
Decision Tree	• Tree structure - Machine tests each node(protocol or packet etc) manually. • Each "leaf" nodes lead to classification	• Easily understandable + fast	• Doesn't really represent complex patterns too well • Picked-up patterns are too broad (overfitting)
Random Forest	• Constructs several "trees"(see above) but each with random subsets of features	• Better generalization • Higher accuracy	• Slower
Artificial Neural Network	• Already went into much depth above	• Extremely effective at capturing complex and specific patterns especially nonlinear patterns • Accurate predictions	• Complex • Relies on a large amount of LABELED data

---

Discussion

Evidence from all of my research documents are both the exponential increase of cybersecurity threats in the digital cyberspace, and the realization that traditional IDS aren't as effective in the modern digital landscape. The recent advancements in Artificial Intelligence and specifically Machine Learning has led to the possibility of a Machine-Learning powered cybersecurity system that could be used to detect cybersecurity threats in real-time, as opposed to relying on a simple signature-recognition system that is prone to zero-day threats.

Of most interest to me was the Cornell University study that tested a Machine Learning model on detecting DDOS attacks on a network. This demonstrated the viability of an in-depth detection system

that worked in real-time. While its focus was mainly on DDOS attacks, I can most likely assert that given a proper database, it could apply to other threats such as viruses and worms. The Machine Learning capabilities demonstrated from this study is something that I'd like to somewhat replicate. I also plan on conducting a benchmark study to compare ANN's with other models that I discussed previously.

Conclusion

The literature that was studied presented a good possibility for the usage of Artificial Neural Networks in Traffic Detection. While it is clear traditional Threat Detection Systems are no longer viable in today's world, Machine Learning Systems - especially Neural Networks - provide a more complex approach that can properly identify patterns in traffic data. However further experimentation and trials through benchmark studies would be necessary to make a definite conclusion about supervised methods.
