Machine Learning Solutions for Cyber Security Challenges

Tarun Tanmay
MBATech - Computer Science
Mukesh Patel School of Technology, Management, and Engineering
Mumbai, India
tanmaysingh0825@gmail.com

Jai Verma
MBATech - Computer Science
Mukesh Patel School of Technology, Management, and Engineering
Mumbai, India
jai.ver.2607@gmail.com

Abstract: Machine Learning has amassed large popularity for itself as it is an emerging field of the present time. It depends on the algorithm teaching itself how to improve in a given task with the adjustment of given weights in the algorithm. While it has already been applied to various domains of the modern world, machine learning is yet to make a big impact in the field of cybersecurity. The need for machine learning in cybersecurity in today's scenario has pushed specialists to start developing neural-network powered algorithms for the same. However, these techniques aren't foolproof yet and have limitations that still restrict us from depending on the human resource for cybersecurity. This paper aims to provide the reader with these limitations and possible solutions that can be used to overcome them. After conducting thorough research and completely analyzing the problems related to machine learning in cybersecurity, we have proposed several solutions in this aspect.

Keywords: Machine Learning, Cyber Security, Neural Networks

INTRODUCTION

Machine learning is a very appealing concept in the existing world of growing computing technologies. This has made it very easy for various domains to automate processes to reduce costs and increase productivity. The domains which have been successful so far include applications of computer vision, natural language processing, and analyzing heaps of data to scan for particular information. While these domains are currently limited to medical fields, social media, marketing, productivity, gaming, and utility tasks, cybersecurity is another field where machine learning has entered. However, machine learning is still not as developed in this field as it is in the former ones to allow for complete automation better than traditional algorithms and human operators. This is currently restricted to finding basic threats in systems and identifying whether a given action could result in a vulnerability or not. However, the lack of advancements still leaves automated cybersecurity as a fairly distant advancement for most people in the current scenario.

This study also aims at delivering the reader with some of the conclusions from extensive research to correctly convey the applications and limitations of machine learning in cybersecurity. While we are not going to focus on the commercial and consumer level applications of machine learning for cybersecurity, we are going to include several modes of cybersecurity analysis to represent a conclusive test for the applications of machine learning for it. We start this with an analysis of where machine learning is currently being used in the field of cybersecurity. In this case, it is mostly restricted to the detection of phishing and spam email, along with the real-time detection of malware and intrusions in a system. We also cover a general analysis of the complexity of ML architectures in the field of cybersecurity that is caused by a lack of available data and time for the training of these machine learning algorithms.

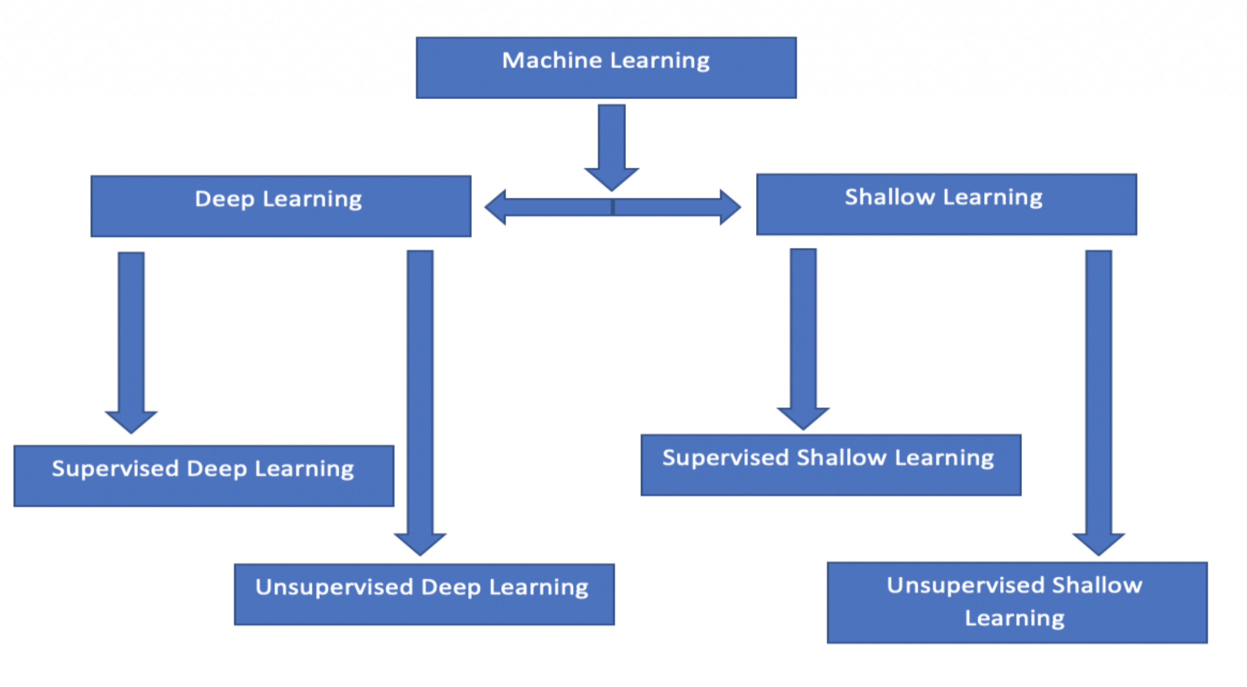
The following section includes the literature that can be used to further understand the applications of neural networks and other machine learning applications in the field of cybersecurity, along with the limitations that they face.

LITERATURE

1.CLASSIFICATIONS OF MACHINE LEARNING ALGORITHMS RELATING TO CYBERSECURITY

Machine learning is a huge field that includes an array of various subjects that are continuously being evolved and modified without much human interaction. Since there are debates about the classifications used here, we cannot fully use the given classifications as a universal standard. However, since the given taxonomy seems to be the most popular choice among specialists, we have chosen to use it to aid in our explanation of the differences between the various methods that can be applied for the detection of cyber threats. The given taxonomy is specifically made for cybersecurity experts and may fall short in pleasing the experts in the machine learning and AI domains.

However, it is derived from those domains and thus shall be used to represent their causes without causing a rift between different experts reading this. Please refer to the below figure for the classifications:


Machine Learning
Deep Learning
Shallow Learning
Supervised Deep Learning
Unsupervised Deep Learning
Supervised Shallow Learning
Unsupervised Shallow Learning

The given figure can differentiate between the broadest forms of machine learning techniques. This enables the reader to finely choose which learning technique would most suit their cause. Next, we have provided a deeper look at the algorithms that are used in these machine learning methods:

Deep Learning

While all of the given deep learning algorithms are founded upon the base of Deep Neural Networks, they each can provide a different approach which can be beneficial or redundant based on the scenario and requirement of the system:

Supervised Deep Learning

Fully Connected Feedforward Deep Neural Networks

These are the variants of Deep Neural Networks that ensure that each neuron is connected to all the neurons of the previous layer. This allows the network to provide a flexible solution for the given problems, though they need higher computational power and therefore incur higher costs.

Convolutional Feedforward Deep Neural Networks

These are the variants of Deep Neural Networks that ensure that each neuron is connected to all the neurons of the previous layer. This allows the network to provide a flexible solution for the given problems, though they need higher computational power and therefore incur higher costs.

Recurrent Deep Neural Networks

These have neurons capable of sending their output to the previous layers, which makes them harder to train. However, they do excel in tasks such as sequence generation.

Unsupervised Deep Learning

Deep Belief Networks

These are neural networks that are modeled after Restricted Boltzmann Machines which have no output layer. These can be used for pre-training tasks and they can themselves be trained with the help of un-labeled datasets.

Stacked Autoencoders

These neural networks are made of a series of autoencoders which have the same number of input and output neurons. These also excel at pre-training tasks, however, do achieve better results than DBNs on smaller datasets.

Shallow Learning

Supervised Shallow Learning

Naïve Bayes

These are probability-based classifier algorithms that assume that the input dataset has independent features. These are scalable and do not require a large training time to start producing accurate results.

Logistic Regression

These are classifiers based on a discriminative model. Similar to the NB algorithm, these can make the independency assumption of all the input features. The performance of these is greatly dependent on the size of the datasets during training.

Support Vector Machines

These are non-probabilistic classifiers that can map data samples in feature spaces to maximize the distance between the categories. These are not able to make assumptions based on the input data and also perform poorly in the case of multi-class classifications. They have limited scalability which mostly restricts them to be used as binary classifiers.

Random Forest

This is a set of decision trees that can consider the individual output of each tree and use this to provide a conclusive response. Each tree works as a conditional classifier and the data goes from the top to bottom while being checked against various conditions at each node. These are excellent methods for large datasets and can take care of multiclass problems.

Hidden Markov Models

These are formed as a set of states that produce outputs based on different probabilistic weights. The goal of these is to determine the sequence of states that resulted in the given outputs. These are excellent for applications where calculating the likelihood of an event is the prime objective. These have been used with labeled datasets in the field of cybersecurity.

K-Nearest Neighbor

These are used to help classify data in the case of multiclass problems. However, their training and usage phases are extremely demanding of computational power.

Shallow Neural Network

This algorithm is based on neural networks that can process elements with components called neurons that can communicate with each other in different layers. SNN includes only neural networks that have a limited number of layers, and they have been utilized for classification tasks in the case of cybersecurity.

Unsupervised Shallow Learning

Clustering

This has group data points that have similar characteristics. This has two approaches – k-means and hierarchal clustering, both of which have limited scalability that offers a flexible solution to be used as a preliminary phase before adopting a supervised algorithm or anomaly inspections.

Association

These aim to identify unknown patterns between data, which makes it suitable for prediction purposes. However, this also tends to produce an excessive output of rules that may not be valid for all stages and may need the help of a human expert.

APPLICATIONS OF MACHINE LEARNING IN CYBERSECURITY

We have considered three separate areas where we feel that machine learning can prove to be beneficial for cybersecurity problems. These are:

Intrusion Detection

This aims to discover several unwarranted activities within a system via an intrusion detection system. Most modern enterprises use Network Intrusion Detection Systems for their networks. While these were initially designed with a comprehensive knowledge of known attacks, they have now evolved to include better detection of newer attacks via machine learning-based classifications. Within the broader scheme of applications, our major focus is on the detection of botnets and Domain Generation Algorithms. Botnets are networks of infected machines that are controlled by attackers to conduct several illegal activities. Despite several research proposals and commercial tools that can address this threat.

Malware Analysis

This is a very important and relevant usage of machine learning since most modern malware can generate its variants which can cause them to appear as different executable files than the original. These are polymorphic and metamorphic features that can defeat rule-based detection approaches. Machine Learning Techniques can be used for the same if they are trained to analyze malware based on certain traits rather than the source code.

Anti-Spam

Spam and phishing have become a very common problem amongst amateur computer users today. This can cause them to send money to illicit users or download malware via email. These days the phishing detection is increasingly made difficult because of the advanced evasion methods used by the attackers to bypass spam filters. Machine learning can improve this as well with the help of training against classifying spam.

PROBLEMS FACED BY MACHINE LEARNING TECHNIQUES IN CYBERSECURITY

Attacker Defender Game

Cybersecurity has contexts that incorporate aspects of game theory which include the age-old attacker-defender game. While human resource-based defense mechanisms can compete with the attackers by rendering attacks unfeasible or unprofitable, the machine learning algorithms will not be able to do so. Even complicated decision trees or Random Forests would not be able to evolve fast enough to predict the attacker’s moves and establish defensive mechanisms in place to counteract them. Any new attack plan would render the security system useless.

Imbalanced Data Sets

For cybersecurity datasets, an imbalanced ratio of 1:10,000 is quite common and this makes it very hard to train the neural networks for DNN strategies. Most machine learning specialists consider this to be an outrageous imbalance since in the machine learning community it maxes out at 1:10.

Tragedy of Metrics

Since there are not a lot of labeled data samples in cybersecurity applications, this can lead to an imbalance of metrics. This means that false positives are going to be quite common. For instance, if you download a game from the internet that contains certain.DLL files, it can be detected by advanced antivirus software as malware since most other malware also has similar.DLL files. This false positive may not be that problematic, however, it can result in downtime in large organizations that would incur costs. On the other hand, a false negative in the data will be disastrous since it can let a dangerous malware slide away just because it is similar to some useful software and unlike most malware.

Domain Adaptation

This includes scenarios where the training model is tested in different distributions than the ones it was built upon. This is important for applications in the cybersecurity world where malware can look completely different from each other or establish variants to make themselves appear different. However, current algorithms are not able to perform adequately in this situation.

Concept Drift

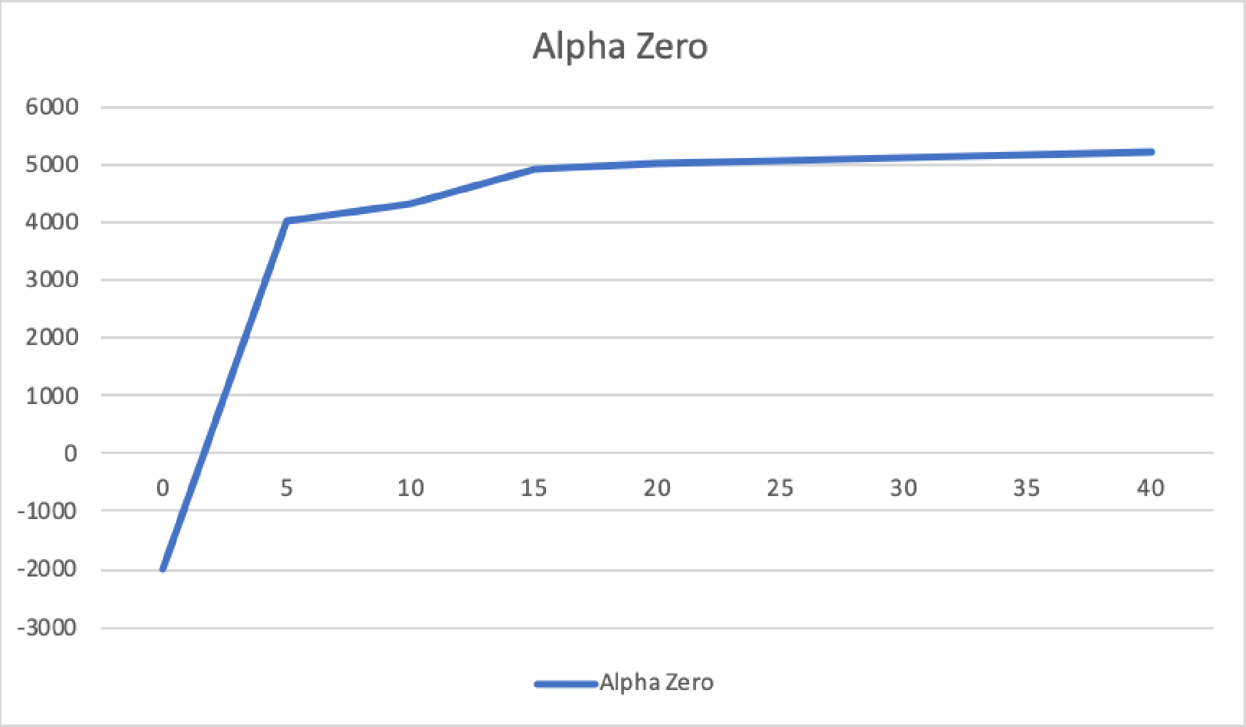
Security threats may cause the system to overload computing resources which can cause a sudden and large degradation in performance shortly afterward since the world of cybersecurity is rapidly changing, and the model has practiced too much in the previous world to be able to evolve again. This can lead to extreme vulnerabilities where attackers launch a decoy attack before their main attack.

A PROPOSED SOLUTION

Furthermore, we have devised some simple concepts that may help in solving the problems that are faced by trying to implement machine learning algorithms in the field of cybersecurity. This is meant to simply try to inform the reader that there are ways to overcome the above problems and not implement these, as it is possible that the practical implementation of these solutions can prove to have more problems associated with it.

AI-Based Attacker Defender Game

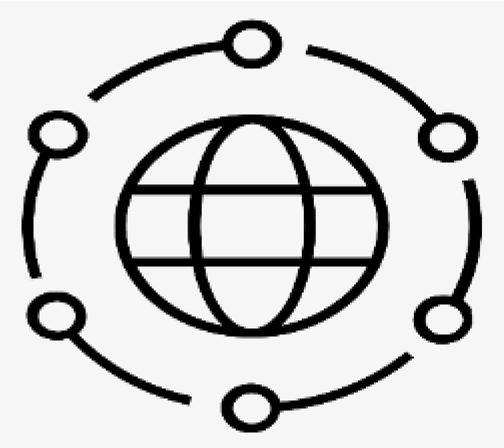
The primary purpose here is to implement an attacker-defender game between two advanced machine learning algorithms. The idea is inspired by Google's implementation of Alpha Go and Alpha Zero, two machine learning powerhouses that have successfully mastered a 4,000-year-old game in less than 45 days. Alpha Zero just learned the entire game by playing against itself in simulated matches for a little over a month without having any prior knowledge of the game except for the rules. This has trained it to win 100% of the matches it plays.



We feel that a similar strategy can be implemented when looking at cybersecurity. The only current problem is to design two algorithms that can use applicable Neural Networks and Monte Carlo tree search systems to create reinforced learning. This would allow them to outsmart each other in the same game that real professionals in cybersecurity use.

Community Support

Since a majority of the problems faced by machine learning in cybersecurity are caused by a lack of proper data sets, we feel it would be helpful to include the community a lot more in the field and get their support. This can help get a large number of labeled data sets and using them to train the algorithms. The current problem in implementing it has been that cybersecurity is not the priority of the machine learning community and vice versa. Getting more researchers to work together on this problem can help in training better algorithms with fewer false metrics.



Quantum Computing

A great way to solve the problem of resources when dealing with cybersecurity problems, it is a good idea to include the field of quantum computing. Shor’s Algorithm, for example, is a major cyber threat with the implementation of quantum computers since it can quickly factorize large numbers and implement this brute force data to hijack network keys.

Quantum computers allow some algorithms that would take thousands of years to complete to run in a matter of hours or even minutes. Now the only problem here is that quantum computing is also an emerging field, so there isn’t much information about it publicly available and the current leaders in the technology are Google and D-wave.

However, it is crucial for the future of cybersecurity to get involved with quantum computing at this early stage and help it grow for improving our security and not destroying it.

Discussion

The research paper has been mostly open-ended, and there are several questions that the reader may think about the information provided here. The primary role of the research paper was to encourage the reader to seek out solutions for existing problems that are faced by machine learning algorithms in cybersecurity. However, it may not be quite possible to implement them without further developments in both fields. Now it is up to the reader to help develop further solutions for the given problem and to encourage others to do the same.

Conclusion

Machine learning approaches are being employed almost everywhere in the world, and have also started to emerge in the field of cybersecurity. It is important in this case to correctly use the right kinds of algorithms and implement them soon. Failing to achieve results with standard ML methods may cause the researchers to believe that the field of cybersecurity is one that cannot be overrun by AI, and hence using more advanced solutions is crucial. We have proposed a taxonomy of the algorithms that can be used, along with some problems faced by the algorithms in their implementation. We have also released some solutions for these problems that can help to overcome them faster.

References

[1] G. Apruzzese, M. Colaganni, L. Ferretti, A. Guido, and M. Marchetti - ‘On the Effectiveness of Machine Learning and Deep Learning for Cyber Security’.

[2] I. Amit, J. Matherly, W. Hewlett, Z. Xu, Y. Meshi, and Y. Weinberger - ’Machine Learning in Cyber-Security – Problems, Challenges, and Data Sets’.

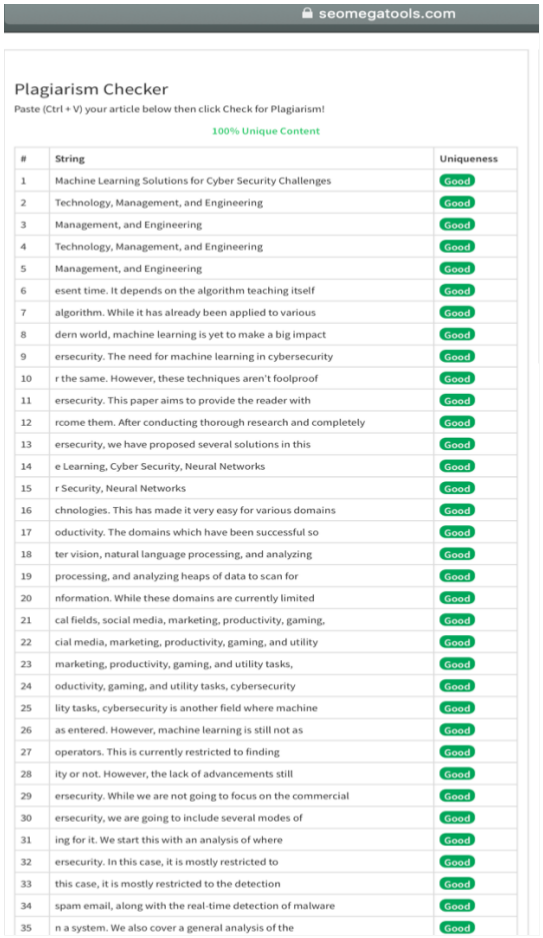
[3] DeepMind Research – ‘AlphaGo Zero: Starting from Scratch.

[4] E. Blanzieri and A. Bryl – ‘A survey of learning techniques of email spam filtering’

[5] Google AI Research – ‘Quantum Supremacy Using a Programmable Superconducting Processor’

Plagiarism Check:

seomegatools.com



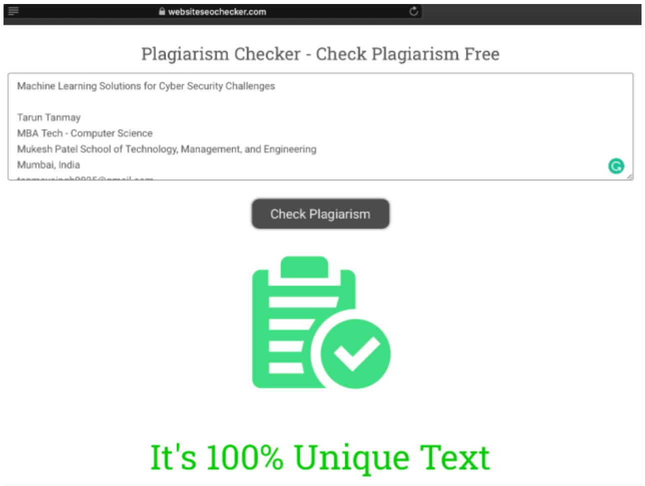
Plagiarism Checker

Paste (Ctrl + V) your article below then click Check for Plagiarism!

100% Unique Content

#	String	Uniqueness
1	Machine Learning Solutions for Cyber Security Challenges	Good
2	Technology, Management, and Engineering	Good
3	Management, and Engineering	Good
4	Technology, Management, and Engineering	Good
5	Management, and Engineering	Good
6	esent time. It depends on the algorithm teaching itself	Good
7	algorithm. While it has already been applied to various	Good
8	dern world, machine learning is yet to make a big impact	Good
9	ersecurity. The need for machine learning in cybersecurity	Good
10	r the same. However, these techniques aren't foolproof	Good
11	ersecurity. This paper aims to provide the reader with	Good
12	rcome them. After conducting thorough research and completely	Good
13	ersecurity, we have proposed several solutions in this	Good
14	e Learning, Cyber Security, Neural Networks	Good
15	r Security, Neural Networks	Good
16	chnologies. This has made it very easy for various domains	Good
17	oductivity. The domains which have been successful so	Good
18	ter vision, natural language processing, and analyzing	Good
19	processing, and analyzing heaps of data to scan for	Good
20	nformation. While these domains are currently limited	Good
21	cal fields, social media, marketing, productivity, gaming,	Good
22	cial media, marketing, productivity, gaming, and utility	Good
23	marketing, productivity, gaming, and utility tasks,	Good
24	oductivity, gaming, and utility tasks, cybersecurity	Good
25	lity tasks, cybersecurity is another field where machine	Good
26	as entered. However, machine learning is still not as	Good
27	operators. This is currently restricted to finding	Good
28	ity or not. However, the lack of advancements still	Good
29	ersecurity. While we are not going to focus on the commercial	Good
30	ersecurity, we are going to include several modes of	Good
31	ing for it. We start this with an analysis of where	Good
32	ersecurity. In this case, it is mostly restricted to	Good
33	this case, it is mostly restricted to the detection	Good
34	spam email, along with the real-time detection of malware	Good
35	n a system. We also cover a general analysis of the	Good

websiteseochecker.com



Plagiarism Checker - Check Plagiarism Free

Machine Learning Solutions for Cyber Security Challenges

Tarun Tanmay
MBA Tech - Computer Science
Mukesh Patel School of Technology, Management, and Engineering
Mumbai, India

It's 100% Unique Text