Machine Learning for Intelligent Data Analysis and Automation in Cybersecurity

Current and Future Prospects

Edited By

Anubhav Bhattacharya

Student at *Techno Main Salt Lake, Kolkata*
Department of Information Technology

Abstract

Due to the digitization and Internet of Things (IoT) revolutions, the present electronic world has a wealth of cybersecurity data. Efficiently resolving cyber anomalies and attacks is becoming a growing concern in today's cyber security industry all over the world. Traditional security solutions are insufficient to address contemporary security issues due to the rapid proliferation of many sorts of cyber-attacks and threats. Utilizing artificial intelligence knowledge, especially machine learning technology, is essential to providing a dynamically enhanced, automated, and up-to-date security system through analyzing security data. In this paper, we provide an extensive view of machine learning algorithms, emphasizing how they can be employed for intelligent data analysis and automation in cybersecurity through their potential to extract valuable insights from cyber data. We also explore a number of potential real-world use cases where data-driven intelligence, automation, and decision-making enable next-generation cyber protection that is more proactive than traditional approaches. The future prospects of machine learning in cybersecurity are eventually emphasized based on our study, along with relevant research directions. Overall, our goal is to explore not only the current state of machine learning and relevant methodologies but also their applicability for future cybersecurity breakthroughs.

Keywords

Cybersecurity · Machine learning · Deep learning · Artificial intelligence · Data-driven decision making · Automation · Cyber analytics · Intelligent systems

Contents

1 Introduction

We live in the digital age, which, like anything else, has its upsides and downsides. The main drawback is the security risk [1]. As more of our sensitive information transfers to the digital arena, security breaches are becoming more common and catastrophic. Cyber-criminals are growing more adept in their attempts to avoid detection, and many newer malware kits are already incorporating new ways to get out of antivirus and other threat detection systems. Cybersecurity, on the other hand, is at a crossroads, and future research efforts should be focused on cyber-attack prediction systems that can foresee important scenarios and consequences, rather than depending on defensive solutions and focusing on mitigation. Systems that are based on a complete,predictive study of cyber risks are required all around the world. The key functionalities in cybersecurity such as *prediction, prevention, identification or detection* as well as corresponding *incident response* should be done *intelligently and automatically*. Artificial intelligence (AI), which is based primarily on *Machine Learning (ML)* [2], is capable of recognizing patterns and predicting future moves based on prior experiences, thereby preventing or detecting potentially malicious activity, which is the primary focus of this study.

ML is one of the most popular current technologies in the fourth industrial revolution (4IR or Industry 4.0) [3] because it allows systems to learn and improve from experience without having to be explicitly programmed. In the cyber security area,machine learning can play a vital role in capturing insights from data. Cybersecurity data can be organized or unstructured, and it can originate from a variety of sources. Intrusion detection, cyber-attack or anomaly detection, phishing or malware detection, zero-day attack prediction, and other intelligent applications can be built by extracting insights from these data. The demand for cybersecurity and protection against cyber anomalies and various sorts of attacks, such as unauthorized access, denial-of-service (DoS), phishing, malware, botnet, spyware, worms,

etc. has risen dramatically in recent days. Thus, real-world cyber applications require *intelligent data analysis tools* and approach capable of extracting insights or meaningful knowledge from data in a timely and intelligent manner. Security researchers believe they can utilize attack pattern recognition or detection methods to provide protection against future attacks.

Learning algorithms can be divided into four categories: supervised, unsupervised, semi-supervised, and reinforcement learning [4]. The nature and quality of the data as well as the effectiveness of the learning algorithms, in general, impact the productivity and efficiency of a machine learning solution. In this paper, we explore various types of machine learning techniques such as classification and regression analysis, security data clustering, rule-based modeling, as well as deep learning approaches, all of which fall within the broad category of machine learning and are capable of building cybersecurity models for different purposes.

1.1 Objective

- To provide a comprehensive understanding of machine learning algorithms that can be applied in cybersecurity for intelligent data analysis and automation.
- To explore the applicability of various machine learning approaches in a variety of real-world scenarios in the context of cybersecurity to allow the next-generation cyber-defense that is more proactive than traditional approaches.
- To emphasize the future prospects of machine learning in cybersecurity, along with relevant research directions.

2 Why Machine Learning in Today's Cybersecurity Research and Applications?

Automation is becoming a key tool for overwhelmed security personnel as today's diverse cyber threats become more widespread, sophisticated, and targeted. Malware, phishing, ransomware, denial-of-service (DoS), zero-day attacks, etc. are common as shown in Fig. 1. This is because most defense measures are not flawless, and many of today's detection approaches rely on an analyst's manual investigation and decision-making to uncover advanced threats, malicious user behavior, and other major associated risks. When it comes to recognizing and predicting specific patterns, machine learning outperforms humans.[4]

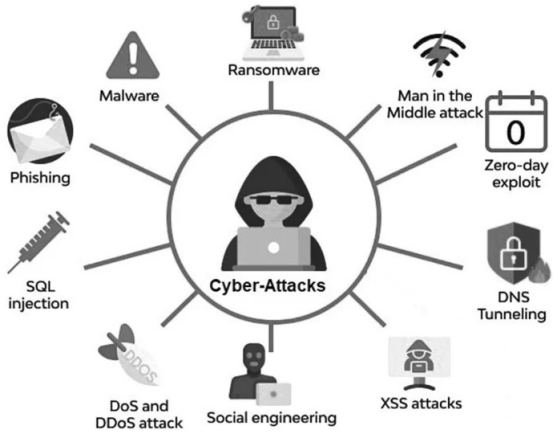

Figure 1: Several common attacks or threats in the context of cybersecurity

In Fig. 2, we have plotted the global statistical impact of machine learning and cybercrime over the previous 5 years, where the x-axis indicates timestamp data and the y-axis represents the equivalent value. We can see from the graph that cybercrime is on the rise all over the world. Thus protecting an information system, especially one that is connected to the Internet, from various cyber-threats, attacks, damage, or unauthorized access is a crucial issue that must be addressed immediately.

Machine learning techniques, with their outstanding learning capabilities from cyber data, can play a vital part in addressing these issues in accordance with today's needs, which is also a popular technology in recent days, as shown in Fig. 2.

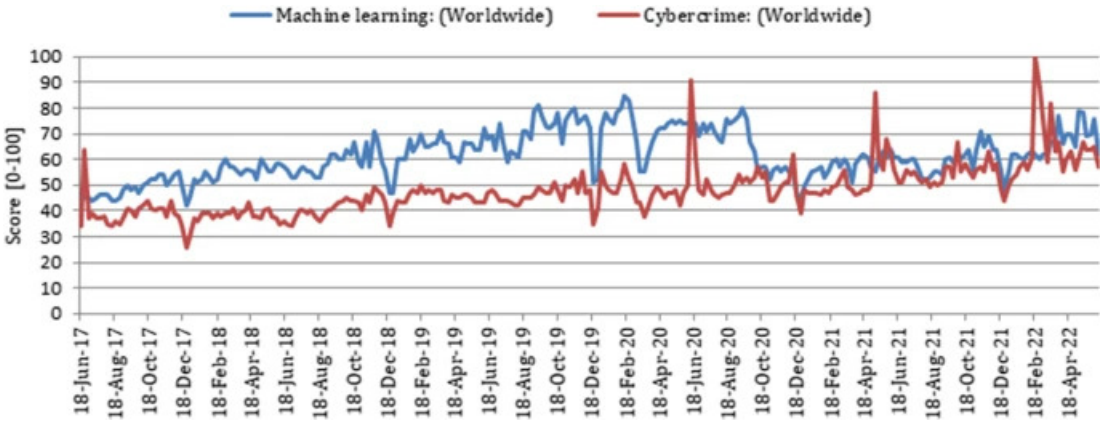

Figure 2: The global statistical impact of machine learning and cybercrime over time, with the x-axis representing the timestamp information and the y-axis representing the equivalent value, on a scale of 0 (min)to 100 (max)

3 Understanding Cybersecurity Data

As machine learning algorithms create models from data, understanding cybersecurity data is essential for intelligent analysis and decision-making. Cybersecurity datasets are often collections of information records that contain a variety of attributes or features, as well as related facts, on which machine learning-based modeling is based. A sample of features from the KDD'99 cup dataset [5] is shown in Fig 1.

For instance, the KDD'99 Cup dataset [6], the most widely used data set including 41 features attributes and a class identification, with attacks divided into four categories: denial of service (DoS), remote-to-local (R2L) intrusions, and user-to-remote (U2R) intrusions, and PROB as well as conven-

tional data. NSL-KDD [12], an updated version of the KDD'99 cup dataset that removes redundant records. Thus a machine learning classification-based security model based on the dataset will not be skewed towards more frequent records.

No.	Features	Types	No.	Features	Types
1	duration	Continuous	22	is_guest_login	Symbolic
2	protocol_type	Symbolic	23	count	Continuous
3	service	Symbolic	24	srv_count	Continuous
4	flag	Symbolic	25	serror_rate	Continuous
5	src_bytes	Continuous	26	srv_serror_rate	Continuous
6	dst_bytes	Continuous	27	rerror_rate	Continuous
7	Land	Symbolic	28	srv_rerror_rate	Continuous
8	wrong_fragment	Continuous	29	same_srv_rate	Continuous
9	urgent	Continuous	30	diff_srv_rate	Continuous
10	hot	Continuous	31	drv_diff_host_rate	Continuous
11	num_failed_logins	Continuous	32	dst_host_count	Continuous
12	logged_in	Symbolic	33	dst_host_srv_count	Continuous
13	num_compromised	Continuous	34	dst_host_same_srv_rate	Continuous
14	root_shell	Continuous	35	dst_host_diff_srv_rate	Continuous
15	su_attempted	Continuous	36	dst_host_same_src_port_rate	Continuous
16	num_root	Continuous	37	dst_host_srv_diff_host_rate	Continuous
17	num_file_creations	Continuous	38	dst_host_serror_rate	Continuous
18	num_shells	Continuous	39	dst_host_srv_serror_rate	Continuous
19	num_access_files	Continuous	40	dst_host_rerror_rate	Continuous
20	num_outbound_cmds	Continuous	41	dst_host_srv_rerror_rate	Continuous
21	is_host_login	Symbolic			

Figure 3: An example of features of KDD'99 cup dataset

4 Machine Learning Tasks and Algorithms in Cybersecurity

Machine learning is typically known as a methodological approach that automates the formation of analytical models, focusing on the use of data and algorithms to mimic the way humans learn while gradually improving accuracy.

A broad structure for a machine learning-based prediction model is shown in Fig. 4, with the model being trained from historical security data contain-

ing benign and malware in phase 1, and the output is generated for new test data in phase 2.

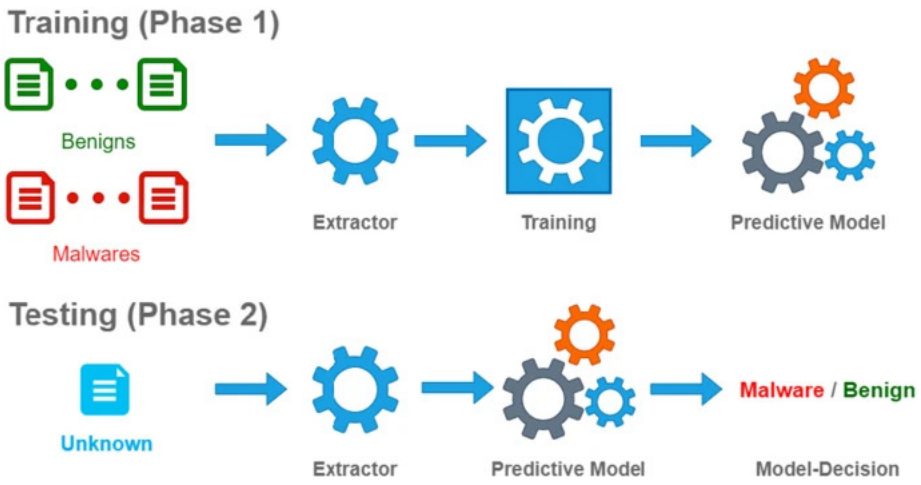

Figure 4: The training and testing phases of a machine learning-based predictive model (i.e., benign or malware)

4.1 Classification and Regression Analysis in Cybersecurity

Both classification and regression approaches are well-known as supervised learning and are frequently employed in the field of machine learning. For instance, an intelligent intrusion detection model for cyber security has been proposed, which is based on the notion of decision trees and takes into account the ranking of security features.

Many classification algorithms have been proposed in the machine learning and data science literature that can be used for intelligent data analysis to solve various real-world issues in the context of cybersecurity. Typically ID3 [7], C4.5 , and CART are well-known DT algorithms in the area of machine learning. K-nearest neighbors, support vector machines, navies Bayes, adaptive boosting, logistic regression, etc. are also popular techniques in

the area. An optimal detection of a phishing attack using SCA-based K-NN has been presented in. To profile abnormal behavior [8] or detect android malware the support vector machine classification technique can be employed[9]. To detect anomalies a naive Bayes based classification model is useful while a logistic regression-based method to detect malicious botnets [1][2].

A regression model, on the other hand, is beneficial for statistically predicting cyberattacks or predicting the impact of an attack, such as worms, viruses, or other malicious software. Because of the enormous dimensionality of cyber security data, regression regularization methods such as Lasso, Ridge, and ElasticNet can improve security breaches analysis.

4.2 Clustering in Cybersecurity

Clustering, which is classified as unsupervised learning, is another common activity in machine learning for processing cybersecurity data. It can cluster or group a set of data points based on measures of similarity and dissimilarity in security data from a variety of sources. Thus clustering may aid in the uncovering of hidden patterns and structures in data, allowing irregularities or breaches to be detected. Clustering data can be done using partition, hierarchy, fuzzy theory, density, and other perspectives [2].

Clustering techniques can help solve a variety of security problems, such as outlier detection, anomaly detection, signature extraction, fraud detection, cyber-attack detection, and so on, by revealing hidden patterns and structures in cybersecurity data and measuring behavioral similarity or dissimilarity. Thus clustering-based unsupervised learning including designing effective algorithms could be a significant topic to explore more for future research in the context of next-generation cybersecurity.

5 Gap Analysis

In the cyber security world, machine learning has become a popular buzzword. As cyber-attacks become more widespread, sophisticated, and targeted, automation is becoming a crucial tool for overloaded security professionals. Cybersecurity is considered a 'zero-tolerance field', meaning that one successful attack results in the security system failing. Cybersecurity, is in a crisis, and future research efforts should be focused on cyber-threat intelligent systems that can predict crucial scenarios and consequences, rather than depending on defensive measures and mitigation. Thus the necessary functions such as "prediction", "prevention", "detection", and "incident response" could be beneficial to a successful and automated cybersecurity system.

Nowadays, deploying good cybersecurity solutions without relying substantially on machine learning is nearly difficult. However, machine learning is challenging to deploy successfully without a thorough, in-depth, and complete approach to the underlying data. Therefore, we should focus more on designing effective machine learning algorithms or data-driven models extracting useful knowledge or security insights as well as data preparing techniques considering real-world raw cyber data,

A future study in the field could include a hybrid learning model, such as an ensemble of methods, updating with an improvement, or designing novel algorithms or models.

6 Conclusion

We have provided a comprehensive view of machine learning techniques for intelligent data analysis and automation in cybersecurity in this paper. For this, we have explored briefly the potentiality of various machine learning techniques to solve practical issues across a range of cyber application fields covered in the paper. The success of a machine learning model depends on how well the data and learning algorithms perform. Prior to the system being able to enable intelligent decision-making and automation, the sophisticated learning algorithms should be trained utilizing real-world cyber data and information particular to the target application, explored in this paper. Finally, we discussed the challenges as well as potential future research directions in the field. Overall, we believe that our study on machine learning-based modeling and security solutions is useful and points in the right direction for further research and application by academics and professionals in the domain of cybersecurity[3].

References

[1] I. H. Sarker, "Smart city data science: Towards data-driven smart cities with open research issues," *Internet of Things*, vol. 19, p. 100528, 2022.

[2] J. M. Tien, "Internet of things, real-time decision making, and artificial intelligence," *Annals of Data Science*, vol. 4, pp. 149–178, 2017.

[3] B. Ślusarczyk, "Industry 4.0: Are we ready?" *Polish Journal of Management Studies*, vol. 17, 2018.

[4] I. H. Sarker, M. H. Furhad, and R. Nowrozy, "Ai-driven cybersecurity: an overview, security intelligence modeling and research directions," *SN Computer Science*, vol. 2, pp. 1–18, 2021.

[5] Q. Wang, Y. Ma, K. Zhao, and Y. Tian, "A comprehensive survey of loss functions in machine learning," *Annals of Data Science*, pp. 1–26, 2020.

[6] M. Gratian, S. Bandi, M. Cukier, J. Dykstra, and A. Ginther, "Correlating human traits and cyber security behavior intentions," *computers & security*, vol. 73, pp. 345–358, 2018.

[7] M. Al-Omari, M. Rawashdeh, F. Qutaishat, M. Alshira'H, and N. Ababneh, "An intelligent tree-based intrusion detection model for cyber security," *Journal of Network and Systems Management*, vol. 29, pp. 1–18, 2021.

[8] I. Rosenberg, A. Shabtai, Y. Elovici, and L. Rokach, "Adversarial machine learning attacks and defense methods in the cyber security domain," *ACM Computing Surveys (CSUR)*, vol. 54, no. 5, pp. 1–36, 2021.

[9] L. Breiman, *Classification and regression trees*. Routledge, 2017.