Cyber Threat Detection Using Machine Learning Techniques: A Performance Evaluation Perspective

Kamran Shaukat, Suhuai Luo, Shan Chen
The University of Newcastle, Australia
Kamran.shaukat@uon.edu.au

Dongxi Liu,
Data61, Commonwealth Scientific and Industrial Research Organization, Australia

Abstract — **The present-day world has become all dependent on cyberspace for every aspect of daily living. The use of cyberspace is rising with each passing day. The world is spending more time on the Internet than ever before. As a result, the risks of cyber threats and cybercrimes are increasing. The term 'cyber threat' is referred to as the illegal activity performed using the Internet. Cybercriminals are changing their techniques with time to pass through the wall of protection. Conventional techniques are not capable of detecting zero-day attacks and sophisticated attacks. Thus far, heaps of machine learning techniques have been developed to detect the cybercrimes and battle against cyber threats. The objective of this research work is to present the evaluation of some of the widely used machine learning techniques used to detect some of the most threatening cyber threats to the cyberspace. Three primary machine learning techniques are mainly investigated, including deep belief network, decision tree and support vector machine. We have presented a brief exploration to gauge the performance of these machine learning techniques in the spam detection, intrusion detection and malware detection based on frequently used and benchmark datasets.**

Keywords— Cyber Threat; Cybercrime; Performance Evaluation; Machine Learning Application; Intrusion Detection System; Malware Detection; Spam Classification

I. INTRODUCTION

The cyberspace refers to the global environment that facilitates the sharing of electronic resources from all over the world. Resources can be an electronic document, audio, video, image, and tweet. The cyberspace incorporates a wide range of components, including the Internet, technically skilled users, system resources, data and untrained users. The cyberspace is providing a global arena to infinitely gain access to information and resources. At present, the cyberspace is playing the leading role in data transfer and information exchange with all its vastly growing losses and gains. After 2017 the cyberspace gained more popularity. Internet usage has risen 81% in developed countries and still growing all over the globe [1]. The elevating cyberspace has also given rise to the risks of cybercrimes and cyber threats.

With the growing range of cyber threats, cyber security has also made a considerable number of enhancements to compete against cybercrimes. The cyber security refers to a set of technologies, technology experts and processes that are used to make safety measures to protect the cyberspace from cybercriminals [2]. There are two main approaches of cyber security, i.e., conventional cyber security and automated cyber security. There are numerous downsides of conventional cyber security which contributes to strengthening cybercrimes, including unqualified users, the weak configuration of system resources and limited access to clean data [3]. The future of cyber security is all about automated cyber security. Advanced and automated cyber security techniques are highly needed. They possess the ability to learn from experience to detect new polymorphic cyberattacks to keep pace with the evolving cybercrimes [4].

The cyber threat is an act in which someone will try or attend to steal the information, violate the integrity rules and harm the computing device or network. Cyber threats include phishing, malware, attack on IoT devices, denial of service attack, spam, intrusion on network or mobile device, financial fraud, ransomware, to name a few [5, 6]. Malware detection, intrusion detection and spam detection are discussed in this paper.

An email that is unwanted or unsolicited is called spam email. Spam emails are mostly used for advertisement or spreading fraudulent material. It occupies the network and computer resources such as the bandwidth of network, memory and wastage of time [7]. Another cyber threat is malware. Malware, as a short for malicious software, is a software that is installed on a computer to disrupt its operation and harm the electronic data. Viruses, worms, ransomware, adware, spyware, malvertising, and Trojan horse are considered as significant types of malware [8]. Malign intrusions over the computer network and devices are another cyber threat to cyberspace. These intrusions are used to identify and scan the vulnerabilities of a network or computer system. An intrusion detection system (IDS) is used to protect against these intrusions. There are three classifications of intrusions, namely, signature/misuse-based, anomaly-based and hybrid [9, 10].

Machine learning (ML) is the most effective and fundamental strategy to compete against cyber threats and overcome the limitations of conventional security systems [11]. Despite having all its charms, machine learning techniques have their constraints and limitations. Machine learning is a subclass of artificial intelligence (AI) [12]. The fascinating quality of machine learning techniques is that machine learning techniques do not need to be explicitly programmed as they can automatically learn from their experience to generate the results [13].

On the strength of all the benefits of machine learning techniques, ML techniques are expanding their scope in almost every area of life, including cyber security [14],

978-1-7281-6840-1/20/$31.00 ©2020 IEEE

medical science [15, 16], educational purposes [17, 18], intrusion detection [19, 20], spam detection [21, 22] and malware detection [23]. Almost all famous machine learning techniques have been applied to detect and classify different cyber threats. Commonly used machine learning techniques are decision tree, random forest, naive Bayes, support vector machine, K-nearest neighbor, deep belief network, artificial neural network, K-mean, to name a few [24, 25]. However, we have considered the decision tree, deep belief network, and support vector machine techniques for this article. We have provided a comparison of machine learning techniques based on frequently used and benchmark datasets.

II. Literature Review

Authors in [26] analyzed the applications of widely used machine learning techniques to protect the cyberspace from cybercriminals. The authors also depicted various obstacles faced during the implementation of machine learning techniques. The work concluded that although the machine learning techniques are expanding various ways to protect cyberspace against cybercriminals, still there is an immense number of advancements needed to protect the classifiers from adversarial attacks. Machine learning classifiers themselves are incredibly vulnerable to cyber threats and adversarial attacks.

Authors in [27] bestowed a brief review of several publications related to the implementation of machine learning models to enhance cyber security. They addressed some commonly faced barriers to machine learning techniques in finding appropriate datasets with most efficient applicability for a specific security problem.

Authors in [28] presented a brief performance comparison of different machine learning techniques, specifically in anomaly detection. They gauged the performance efficiency of feature selection in ML for IDS. They claimed that the convolutional neural network (CNN) classifier is an underused classifier and it could have brought vast advancements in cyber security if it was used to its full potential.

Authors in [29] analyzed the role of various machine learning techniques in spam detection, malware detection and intrusion detection. They claimed that there is no machine learning technique that is not vulnerable to cyberattacks. Every machine learning technique is still struggling to keep a pace with continuously upgrading cybercrimes.

Authors in [30] proposed a novel machine learning technique for spam detection in text messages using content-based features. They concluded that the proposed averaged neural network and content-based feature selection outplayed most of the recent machine learning techniques in terms of accuracy on the same dataset. Authors in [31] stated that the signature-based classification techniques generate results with high error rates when it comes to mobile malware detection. They proposed an image-based deep learning technique for mobile malware detection, aiming to demonstrate the discrimination between the family of malicious attributes and the legitimate attributes by obtaining grey-scale images.

Authors in [32] came up with a statistical semi-supervised machine learning technique for intrusion detection in Android mobile devices. The increase in data traffic will also give rise to cybercrimes. Consequently, to protect Android mobile devices against advanced cybercrimes, more advanced machine learning techniques are needed to be developed to detect malicious activities.

In this paper, we have provided a comprehensive review of widely used machine learning techniques to gauge the performance of machine learning techniques to detect some widely known cybercrimes. We have analyzed three widely used machine learning techniques, namely: decision tree, deep belief network and support vector machine. Most of the review articles only focused on a particular threat. However, we have considered three major cyber threats. An intrusion detection, spam detection and malware detection are considered for this study. We have provided a comprehensive comparison to see the performance of each classifier based on frequently used datasets. We have mentioned the computational complexity of each classifier. The following section will discuss the fundamentals of machine learning, an overview of considered classifiers and evaluation criteria to evaluate the performance of a classifier. The discussion section will discuss cyber threats and provide the performance evaluation in the form of accuracy, recall and precision. Lastly, the conclusion section will conclude the study.

III. Fundamentals Of Machine Learning

Artificial intelligence is a branch of computer science based on simulation of the human brain by an artificial entity to automate a necessary process. Machine learning is a sub-branch of AI. It achieves a specific goal by using the results from experience without explicitly being programmed. Hence machine learning does not require to be fed explicitly with data [33]. There are three sub-branches of machine learning, namely, supervised learning, unsupervised learning and semi-supervised learning. In supervised learning, the targeted class/label is known in advance, whereas the targeted classes are unknown in unsupervised learning. Unsupervised learning divides the data into different clusters based on the similarity between data objects. Semi-supervised learning combines characteristics of both: supervised learning and unsupervised learning.

Decision tree, random forest, naive Bayes, support vector machine, K-nearest neighbour, deep belief network, artificial neural network, K-mean are widely used learning techniques to detect cyber threats. We have considered three techniques that are decision tree, deep belief network, and support vector machine. We have briefly described each technique below.

A deep belief network (DBN) is a complex representation of middle layers of Restricted Boltzmann Machine (RBM). Deep belief network follows a greedy approach. Every layer communicates with the previous layer and the next layer. In each layer of the deep belief network, the nodes do not communicate laterally with other nodes. In a deep belief network, every layer is assigned with both input and output tasks, excluding the first layer and the last layer. The end layer is the classifier layer. The computation complexity of DBN is

O((n+N)k) where k is the number of iterations, n represents the number of records, and N is the number of parameters in DBN [34].

Decision tree (DT) is a supervised machine learning technique. The main components of a decision tree are nodes, paths and leaf nodes. A node can be a root node or an intermediate node. Decision tree follows the if-then rule to find the best suitable root node at each level. Leaf node or terminal node is an ending node. The decision class is denoted by the leaf node [35]. The time complexity of DT is $O(mn^2)$ where n represents the number of instances and m shows the number of attributes [36, 37].

TABLE I. CONFUSION MATRIX

	Predicted as Normal	Predicted as Attack
Actual Labeling as Normal	$T_{Positive}$	$F_{Negative}$
Actual Labeling as Attack	$F_{Positive}$	$T_{Negative}$

Support vector machine (SVM) is another widely used supervised machine learning model. SVM works to find hyperplane with most suitable dataset distribution by classifying the data into two classes on both sides of the hyperplane. Both sides of the hyperplane donate a separate class. The class of every data point depends on the side of the hyperplane it lands. Support vector machine has a high consumption of space and time to handle larger and noisier datasets [25]. The computational complexity of SVM is $O(n^2)$ where n represents the number of instances [38, 39].

A matrix that is used to evaluate the performance of machine learning classifier is called a confusion matrix [40], as depicted in Table 1. $T_{Positive}$ means the number of normal instances that are correctly classified as normal. $T_{Negative}$ means the number of attack instances that are correctly classified as an attack. $F_{Negative}$ means the number of normal instances that are misclassified as an attack. $F_{Positive}$ means the number of attack instances that are misclassified as normal.

Precision

The precision is a percentage of the total number of positive instances classified to the total number of positive instances.

$$\text{Precision}= T_{Positive} / (T_{Positive} + F_{Positive}) \quad (1)$$

Error Rate

The error rate (ERate) is a percentage of the total number of misclassified instances to all instances of the dataset.

$$\text{ERate} = (F_{Positive} + F_{Negative})/ (T_{Negative} + F_{Positive} + F_{Negative} + T_{Positive}) \quad (2)$$

Recall

The recall is a percentage of correctly classified positive instances to the total number of positive instances classified in the dataset.

$$\text{Recall} = T_{Positive} / (T_{Positive} + F_{Negative}) \quad (3)$$

IV. Discussion and Performance Evalaution

There is a wide range of cybercrimes that try to breach the privacy of user's data daily on a computer network or mobile devices. An extensive range of machine learning techniques have been developed to battle against cybercrimes. However those techniques are still lagging a step behind as compared to cybercrimes. In our review, we have mainly focused on the detection of three cardinal cyber threats, namely: IDS, malware detection and spam detection. We have considered three learning models that are decision tree, support vector machine, and deep belief network. Datasets play an important role in completing all the significant tasks as the results are all dependent on the type and size of the dataset. The diversity of the dataset helped to evaluate the performance of the classifier in the training and testing phases. Real-time and diverse datasets produce better results than a customized dataset. In this review, we have considered frequently used and benchmark datasets that are KDD CUP 99 [41], Spambase [42], Twitter dataset [43], Enron [44], NSL-KDD [45], DARPA [46], and malware datasets [47]. We have compared the performance of the machine learning models on detecting these cyber threats.

TABLE II. PERFORMANCE RESULTS OF SPAM DETECTION USING MACHINE LEARNING MODELS

Cyber Threat	Learning Model	Dataset	Reference	Published Year	Sub-Domain	Performance Results		
						Precision	Accuracy	Recall
Spam Detection	Support Vector Machine	Spambase	[48]	2011	Email Spam	93.12 %	96.90 %	95.00 %
			[49]	2015	Email Spam	79.02 %	79.50 %	68.67 %
		Twitter Dataset	[50]	2018	Spam Tweets	92.91 %	93.14 %	93.14 %
			[51]	2015	Spam Tweets		95.20 %	93.60 %
	Decision Tree	Enron	[52]	2016	Email Spam	98.00 %	96.00 %	94.00 %
			[52]	2016	Email Spam	98.00 %	96.00 %	94.00 %
		Spambase	[53]	2014	Email Spam	91.51 %	92.08 %	88.08 %
			[54]	2014	Email Spam	-	94.27 %	91.02 %
	DBN	Enron	[55]	2016	Email Spam	96.49 %	95.86 %	95.61 %
			[56]	2016	Email Spam	98.39 %	97.50 %	98.02 %
		Spambase	[57]	2007	Email Spam	94.94 %	97.43 %	96.47 %
			[58]	2018	Email Spam	96.00 %	89.20 %	-

We have taken accuracy, recall and precision as evaluation factors to measure the performance of classification models. Table 2, Table 3, and Table 4 present the performance of three learning models for spam detection, malware detection, and intrusion detection, respectively. Cyber Threat and Learning Model columns are self-explanatory. Dataset column shows the frequently used and benchmark dataset for each particular threat. Reference column depicts the citation of specific paper that shows the evaluation results. Values for the sub-domain column is different for each cyber threat. Performance results column shows the performance results of each cited article. Following sub-sections will present the discussion on each cyber threat.

A. Spam Detection

Spam is a threat to computer and network resources. It is a term used for an unwanted message. Spam can be in different mediums. It can be in the form of text messages, images and videos on mobile devices [59]. Spam tweets and spam emails are the mediums that are mostly used over the computing devices and network. Spam messages consume a lot of network resources, such as bandwidth. Spam emails in the form of unnecessary advertisements consume a lot of time. Machine learning techniques have been applied in the literature to distinguish between a genuine email and a spam email, as shown in Table 2. SVM and DT have shown a good accuracy of 96.90 % [48]. However, DBN has outperformed with a precision value of 98.39 % using Enron dataset [56]. DBN also outperformed in terms of recall and precision over SVM and DT. On Spambase dataset, SVM performed better than DT with an accuracy of 96.90 % [48]. Using Enron dataset, the decision tree has shown better precision than SVM and similar precision to DBN [52]. It is apparent from Table 2 that DBN has performed better than other learning models for these particular datasets. Based on the above evaluation metrics, the authors recommend using DBN for spam detection.

B. Intrusion Detection

Malign intrusions over the computer network and devices are another cyber threat to cyberspace. These intrusions are used to identify the vulnerabilities of a network [60]. Intrusions identify the weakness within a computer system for further attacks. An intrusion detection system is used to protect against these intrusions. There are three classifications of intrusions, namely, signature/misuse-based, anomaly-based and hybrid [61]. The intrusions can be detected on the network or a host computer. Conventional techniques are unable to cope with the pace to detect intrusions. Commonly used datasets are DARPA and KDD versions. However, these datasets are older for more than fifteen years. Table 3 presents the evaluation results of intrusion detection. DBN performed better than SVM and DT in terms of accuracy. DBN has shown better accuracy results of 96.70 % using NSL-KDD dataset [62].

TABLE III. PERFORMANCE RESULTS OF INTRUSION DETECTION SYSTEM USING MACHINE LEARNING MODELS

Cyber Threat	Learning Model	Dataset	Reference	Published Year	Sub-Domain	Performance Results		
						Precision	Accuracy	Recall
Intrusion Detection	Support Vector Machine	NSL-KDD	[63]	2019	Anomaly-Based	-	89.70 %	-
			[41]	2014	Hybrid-Based	74.00 %	82.37 %	82.00 %
		DARPA	[64]	2007	Hybrid-Based	-	69.80 %	-
			[65]	2014	Anomaly-Based	-	95.11 %	-
	Decision Tree	KDD	[66]	2018	Misuse-Based	-	99.96 %	-
			[67]	2017	Hybrid-Based	-	86.29 %	78.00 %
		NSL-KDD	[68]	2019	Hybrid-Based	-	93.40 %	-
			[69]	2017	Hybrid-Based	91.15 %	90.30 %	90.31 %
	DBN	KDD	[61]	2015	Anomaly-Based	-	97.50 %	-
		NSL-KDD	[62]	2015	Hybrid-Based	97.90 %	96.70 %	-
			[70]	2017	Anomaly-Based	88.60 %	90.40 %	95.30 %

TABLE IV. PERFORMANCE RESULTS OF MALWARE DETECTION USING MACHINE LEARNING MODELS

Cyber Threat	Learning Model	Dataset	Reference	Published Year	Sub-Domain	Performance Results		
						Precision	Accuracy	Recall
Malware Detection	Support Vector Machine	Malware Dataset	[71]	2017	Static	-	94.37 %	-
			[72]	2013	Dynamic	-	95.00 %	-
		Enron	[73]	2015	Dynamic	-	97.10 %	-
			[52]	2016	Static	84.74 %	91.00 %	100 %
	Decision Tree	Custom	[74]	2016	Static	99.40 %	99.90 %	-
		Malware Dataset	[75]	2017	Static	-	84.70 %	-
			[76]	2014	Static	97.90 %	-	96.70 %
	DBN	Custom	[77]	2016	Dynamic	78.08 %	71.00 %	59.09 %
			[77]	2016	Static	83.00 %	89.03 %	98.18 %
		KDD CUP99	[77]	2016	Hybrid	95.77 %	96.76 %	97.84 %
			[78]	2015	Hybrid	-	91.40 %	95.34 %

However, the decision tree has shown outstanding accuracy of 99.96 % and better than DBN and SVM using KDD dataset [66]. The decision tree has shown the best efficiency among the learning classifiers of 99.96 % regardless of the dataset [66]. DBN has reported the best recall and precision values of 95.30 % and 97.90 %, respectively [62, 70]. Based on the considered articles, the decision tree is recommended as the best learning classifier for intrusion detection, as depicted in Table 3.

C. *Malware detection*

Malware, short for malicious software, is a software that is installed on a computer to disrupt its operation and harm the electronic data. Viruses, worms, ransomware, adware, spyware, malvertising, and Trojan horse are considered as significant types of malware [79]. Malware interrupts the normal flow of computer operations. With a growing pace of usage of computing and mobile devices, the cybercriminal is finding it easy to compromise the integrity of data. Malware also disrupts the availability of computer and network resources. Machine learning techniques are being used to detect malware. The performance of each learning classifier is depicted in Table 4. Static detection is a sub-domain of malware detection in which applications are tested for malware without executing them. However, in dynamic detection, the applications or software are tested by executing them. Hybrid detection is a mixture of both static and dynamic detection [80]. The decision tree has shown overall best accuracy of 99.90 % on custom data collected by the author [74]. However, on a malware dataset, SVM performed better than decision tree in terms of accuracy. SVM reported the best recall value of 100% [52]. SVM is recommended based upon the cited papers to detect and classify applications from malware.

V. Conclusion

Cyber threats are increasing at a growing pace. The conventional security techniques are not capable enough of coping with these threats. Machine learning techniques are being applied to overcome the limitations of conventional security systems. Machine learning techniques are playing their role at both ends: at defender-end and attacker-end. We have presented a performance comparison of three learning models to detect and classify the intrusion, spam and malware. We have considered frequently used and benchmark datasets to compare the evaluation results in terms of recall, precision, and accuracy. In the previous section, we have discussed and concluded that we cannot recommend a particular learning technique for every cyber threat detection. Different learning models are being used for specific different cyber threats. On the other hand, there is a vast number of authors who have worked to highlight the constraints faced by machine learning techniques. We have observed and suggested that there is a dare need of latest benchmark dataset to test the latest advancement in the field of machine learning for cyber threat detection. Available datasets lack in terms of diversity and sophisticated attacks and contain missing values. There is a need for specific and customized learning models specifically designed for security purposes. In future, we will focus on analyzing more learning techniques for cyber threat detection.

References

[1] "ICT Facts and Figures 2017." Telecommunication Development Bureau,International Telecommunication Union (ITU), Technical Report. https://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx (accessed October 09, 2019).

[2] "What is Cyber-Security?" https://www.kaspersky.com.au/resource-center/definitions/what-is-cyber-security (accessed January 11, 2020).

[3] F. Farahmand, S. B. Navathe, P. H. Enslow, and G. P. Sharp, "Managing vulnerabilities of information systems to security incidents," in *Proceedings of the 5th international conference on Electronic commerce*, 2003: ACM, pp. 348-354.

[4] P. Szor, *The Art of Computer Virus Research and Defense: ART COMP VIRUS RES DEFENSE _p1*. Pearson Education, 2005.

[5] M. Jump, "Fighting Cyberthreats with Technology Solutions," *Biomedical instrumentation & technology,* vol. 53, no. 1, pp. 38-43, 2019.

[6] N. Kostyuk and C. Wayne, "Communicating Cybersecurity: Citizen Risk Perception of Cyber Threats," 2019.

[7] A. K. Jain, D. Goel, S. Agarwal, Y. Singh, and G. Bajaj, "Predicting Spam Messages Using Back Propagation Neural Network," *Wireless Personal Communications,* vol. 110, no. 1, pp. 403-422, 2020.

[8] "Malware Types and Classifications." https://www.lastline.com/blog/malware-types-and-classifications/ (accessed April 18,2020).

[9] N. Sultana, N. Chilamkurti, W. Peng, and R. Alhadad, "Survey on SDN based network intrusion detection system using machine learning approaches," *Peer-to-Peer Networking and Applications,* vol. 12, no. 2, pp. 493-501, 2019.

[10] M. Pradhan, C. K. Nayak, and S. K. Pradhan, "Intrusion Detection System (IDS) and Their Types," in *Securing the Internet of Things: Concepts, Methodologies, Tools, and Applications*: IGI Global, 2020, pp. 481-497.

[11] I. Firdausi, A. Erwin, and A. S. Nugroho, "Analysis of machine learning techniques used in behavior-based malware detection," in *2010 second international conference on advances in computing, control, and telecommunication technologies*, 2010: IEEE, pp. 201-203.

[12] A. V. Joshi, *Machine Learning and Artificial Intelligence*. Springer, 2020.

[13] D. Michie, D. J. Spiegelhalter, and C. Taylor, "Machine learning," *Neural and Statistical Classification,* vol. 13, 1994.

[14] K. Shaukat, A. Rubab, I. Shehzadi, and R. Iqbal, "A Socio-Technological analysis of Cyber Crime and Cyber Security in Pakistan," *Transylvanian Review,* vol. 1, no. 3, 2017.

[15] K. Shaukat, N. Masood, A. B. Shafaat, K. Jabbar, H. Shabbir, and S. Shabbir, "Dengue fever in perspective of clustering algorithms," *arXiv preprint arXiv:1511.07353,* 2015.

[16] K. Shaukat, N. Masood, S. Mehreen, and U. Azmeen, "Dengue fever prediction: A data mining problem," *Journal of Data Mining in Genomics & Proteomics,* vol. 2015, 2015.

[17] K. Shaukat, I. Nawaz, and S. Zaheer, *Students Performance: A Data Mining Perspective*. LAP Lambert Academic Publishing, 2017.

[18] K. Shaukat, I. Nawaz, S. Aslam, S. Zaheer, and U. Shaukat, "Student's performance in the context of data mining," in *2016 19th International Multi-Topic Conference (INMIC)*, 2016: IEEE, pp. 1-8.

[19] S. Dey, Q. Ye, and S. Sampalli, "A machine learning based intrusion detection scheme for data fusion in mobile clouds involving heterogeneous client networks," *Information Fusion,* vol. 49, pp. 205-215, 2019.

[20] B. Geluvaraj, P. Satwik, and T. A. Kumar, "The future of cybersecurity: Major role of artificial intelligence, machine learning, and deep learning in cyberspace," in *International Conference on Computer Networks and Communication Technologies*, 2019: Springer, pp. 739-747.

[21] A. A. Alurkar *et al.*, "A Comparative Analysis and Discussion of Email Spam Classification Methods Using Machine Learning Techniques," *Applied Machine Learning for Smart Data Analysis,* p. 185, 2019.

[22] E. G. Dada, J. S. Bassi, H. Chiroma, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems," *Heliyon,* vol. 5, no. 6, p. e01802, 2019.

[23] P. Jain, "Machine Learning versus Deep Learning for Malware Detection," 2019.

[24] P. Thiyagarajan, "A Review on Cyber Security Mechanisms Using Machine and Deep Learning Algorithms," in *Handbook of Research on Machine and Deep Learning Applications for Cyber Security*: IGI Global, 2020, pp. 23-41.

[25] S. S. Iyer and S. Rajagopal, "Applications of Machine Learning in Cyber Security Domain," in *Handbook of Research on Machine and Deep Learning Applications for Cyber Security*: IGI Global, 2020, pp. 64-82.

[26] V. Ford and A. Siraj, "Applications of machine learning in cyber security," in *Proceedings of the 27th International Conference on Computer Applications in Industry and Engineering*, 2014.

[27] H. Jiang, J. Nagra, and P. Ahammad, "Sok: Applying machine learning in security-a survey," *arXiv preprint arXiv:1611.03186,* 2016.

[28] E. Hodo, X. Bellekens, A. Hamilton, C. Tachtatzis, and R. Atkinson, "Shallow and deep networks intrusion detection system: A taxonomy and survey," *arXiv preprint arXiv:1701.02145,* 2017.

[29] G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido, and M. Marchetti, "On the effectiveness of machine and deep learning for cyber security," in *2018 10th International Conference on Cyber Conflict (CyCon)*, 2018: IEEE, pp. 371-390.

[30] S. Sheikhi, M. Kheirabadi, and A. Bazzazi, "An Effective Model for SMS Spam Detection Using Content-based Features and Averaged Neural Network," *International Journal of Engineering,* vol. 33, no. 2, pp. 221-228, 2020.

[31] F. Mercaldo and A. Santone, "Deep learning for image-based mobile malware detection," *Journal of Computer Virology and Hacking Techniques,* pp. 1-15, 2020.

[32] J. Ribeiro, F. B. Saghezchi, G. Mantas, J. Rodriguez, S. J. Shepherd, and R. A. Abd-Alhameed, "An autonomous host-based intrusion detection system for android mobile devices," *Mobile Networks and Applications,* vol. 25, no. 1, pp. 164-172, 2020.

[33] C. Chen *et al.*, "A performance evaluation of machine learning-based streaming spam tweets detection," *IEEE Transactions on Computational social systems,* vol. 2, no. 3, pp. 65-76, 2015.

[34] Z. Chen, S. Liu, K. Jiang, H. Xu, and X. Cheng, "A data imputation method based on deep belief network," in *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, 2015: IEEE, pp. 1238-1243.

[35] D. M. Farid, N. Harbi, and M. Z. Rahman, "Combining naive bayes and decision tree for adaptive intrusion detection," *arXiv preprint arXiv:1005.4496,* 2010.

[36] Q. J. Ross, "C4. 5: programs for machine learning," *San Mateo, CA,* 1993.

[37] P. S. Oliveto, J. He, and X. Yao, "Time complexity of evolutionary algorithms for combinatorial optimization: A decade of results," *International Journal of Automation and Computing,* vol. 4, no. 3, pp. 281-293, 2007.

[38] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery,* vol. 2, no. 2, pp. 121-167, 1998.

[39] G. D. Forney, "The viterbi algorithm," *Proceedings of the IEEE,* vol. 61, no. 3, pp. 268-278, 1973.

[40] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, "An improved method to construct basic probability assignment based on the confusion matrix for classification problem," *Information Sciences,* vol. 340, pp. 250-261, 2016.

[41] M. S. Pervez and D. M. Farid, "Feature selection and intrusion classification in NSL-KDD cup 99 dataset employing SVMs," in *The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014)*, 2014: IEEE, pp. 1-6.

[42] "Spambase Dataset. Center for Machine Learning and Intelligent Systems at UC Irvine." https://archive.ics.uci.edu/ml/datasets/Spambase (accessed January 31, 2020).

[43] D. Gunawan, R. F. Rahmat, A. Putra, and M. F. Pasha, "Filtering Spam Text Messages by Using Twitter-LDA Algorithm," in *2018 IEEE International Conference on Communication, Networks and Satellite (Comnetsat)*, 2018: IEEE, pp. 1-6.

[44] B. Klimt and Y. Yang, "Introducing the Enron corpus," in *CEAS*, 2004.

[45] B. Ingre and A. Yadav, "Performance analysis of NSL-KDD dataset using ANN," in *2015 International Conference on Signal Processing and Communication Engineering Systems*, 2015: IEEE, pp. 92-96.

[46] A. Chahal and R. Nagpal, "Performance of Snort on Darpa Dataset and Diferent False Alert Reduction Techniques," in *3rd International Conference on Electrical, Electronics, Engineering Trends, Communication, Optimization and Sciences (EEECOS).*

[47] H. Kim, T. Cho, G.-J. Ahn, and J. H. Yi, "Risk assessment of mobile applications based on machine learned malware dataset," *Multimedia Tools and Applications,* vol. 77, no. 4, pp. 5027-5042, 2018.

[48] W. Awad, S. J. I. J. o. C. S. ELseuofi, and I. Technology, "Machine learning methods for spam e-mail classification," vol. 3, no. 1, pp. 173-184, 2011.

[49] R. Karthika and P. J. W. T. C. Visalakshi, "A hybrid ACO based feature selection method for email spam classification," vol. 14, pp. 171-177, 2015.

[50] G. Jain, M. Sharma, and B. J. I. J. o. K. D. i. B. Agarwal, "Spam detection on social media using semantic convolutional neural network," vol. 8, no. 1, pp. 12-26, 2018.

[51] C. Chen *et al.*, "A performance evaluation of machine learning-based streaming spam tweets detection," vol. 2, no. 3, pp. 65-76, 2015.

[52] Z. Khan and U. Qamar, "Text Mining Approach to Detect Spam in Emails," in *The International Conference on Innovations in Intelligent Systems and Computing Technologies (ICIISCT2016)*, 2016, p. 45.

[53] S. A. Saab, N. Mitri, and M. Awad, "Ham or spam? A comparative study for some content-based classification algorithms for email filtering," in *MELECON 2014-2014 17th IEEE Mediterranean Electrotechnical Conference*, 2014: IEEE, pp. 339-343.

[54] Y. Zhang, S. Wang, P. Phillips, and G. J. K.-B. S. Ji, "Binary PSO with mutation operator for feature selection using decision tree applied to spam detection," vol. 64, pp. 22-31, 2014.

[55] A. Tyagi, "Content Based Spam Classification-A Deep Learning Approach," University of Calgary, 2016.

[56] I. J. Alkaht and B. J. I. R. C. S. Al-Khatib, "Filtering SPAM Using Several Stages Neural Networks," vol. 11, p. 2, 2016.

[57] G. Tzortzis and A. Likas, "Deep belief networks for spam filtering," in *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, 2007, vol. 2: IEEE, pp. 306-309.

[58] Y. Rizk, N. Hajj, N. Mitri, M. J. A. C. Awad, and Informatics, "Deep belief networks and cortical algorithms: A comparative study for supervised classification," 2018.

[59] A. Sharaff, N. K. Nagwani, and A. Dhadse, "Comparative study of classification algorithms for spam email detection," in *Emerging research in computing, information, communication and applications*: Springer, 2016, pp. 237-244.

[60] C. Yin, Y. Zhu, J. Fei, and X. He, "A deep learning approach for intrusion detection using recurrent neural networks," *Ieee Access,* vol. 5, pp. 21954-21961, 2017.

[61] M. Z. Alom, V. Bontupalli, and T. M. Taha, "Intrusion detection using deep belief networks," in *2015 National Aerospace and Electronics Conference (NAECON)*, 2015: IEEE, pp. 339-344.

[62] S. Jo, H. Sung, and B. Ahn, "A comparative study on the performance of intrusion detection using decision tree and artificial neural network models," *Journal of the Korea Society of Digital Industry and Information Management,* vol. 11, no. 4, pp. 33-45, 2015.

[63] J. Lee, J. Kim, I. Kim, and K. Han, "Cyber Threat Detection Based on Artificial Neural Networks Using Event Profiles," *IEEE Access,* vol. 7, pp. 165607-165626, 2019.

[64] L. Khan, M. Awad, and B. Thuraisingham, "A new intrusion detection system using support vector machines and hierarchical clustering," *The VLDB journal,* vol. 16, no. 4, pp. 507-521, 2007.

[65] R. Kokila, S. T. Selvi, and K. Govindarajan, "DDoS detection and analysis in SDN-based environment using support vector machine classifier," in *2014 Sixth International Conference on Advanced Computing (ICoAC)*, 2014: IEEE, pp. 205-210.

[66] P. Mishra, V. Varadharajan, U. Tupakula, E. S. J. I. C. S. Pilli, and Tutorials, "A detailed investigation and analysis of using machine learning techniques for intrusion detection," vol. 21, no. 1, pp. 686-728, 2018.

[67] J. Kevric, S. Jukic, A. J. N. C. Subasi, and Applications, "An effective combining classifier approach using tree algorithms for network intrusion detection," vol. 28, no. 1, pp. 1051-1058, 2017.

[68] A. Ahmim, L. Maglaras, M. A. Ferrag, M. Derdour, and H. Janicke, "A novel hierarchical intrusion detection system based on decision tree and rules-based models," in *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 2019: IEEE, pp. 228-233.

[69] B. Ingre, A. Yadav, and A. K. Soni, "Decision tree based intrusion detection system for NSL-KDD dataset," in *International Conference on Information and Communication Technology for Intelligent Systems*, 2017: Springer, pp. 207-218.

[70] D. Kwon, H. Kim, J. Kim, S. C. Suh, I. Kim, and K. J. Kim, "A survey of deep learning-based network anomaly detection," *Cluster Computing,* pp. 1-13, 2017.

[71] Y. Cheng, W. Fan, W. Huang, and J. An, "A Shellcode Detection Method Based on Full Native API Sequence and Support Vector Machine," in *IOP Conference Series: Materials Science and Engineering*, 2017, vol. 242, no. 1: IOP Publishing, p. 012124.

[72] A. Mohaisen and O. Alrawi, "Unveiling zeus: automated classification of malware samples," in *Proceedings of the 22nd International Conference on World Wide Web*, 2013: ACM, pp. 829-832.

[73] P. Shijo and A. J. P. C. S. Salim, "Integrated static and dynamic analysis for malware detection," vol. 46, pp. 804-811, 2015.

[74] Q. Jamil and M. A. Shah, "Analysis of machine learning solutions to detect malware in android," in *2016 Sixth International Conference on Innovative Computing Technology (INTECH)*, 2016: IEEE, pp. 226-232.

[75] D. Moon, H. Im, I. Kim, and J. H. Park, "DTB-IDS: an intrusion detection system based on decision tree using behavior analysis for preventing APT attacks," *The Journal of supercomputing,* vol. 73, no. 7, pp. 2881-2895, 2017.

[76] Z. Salehi, A. Sami, M. J. C. F. Ghiasi, and Security, "Using feature generation from API calls for malware detection," vol. 2014, no. 9, pp. 9-18, 2014.

[77] Z. Yuan, Y. Lu, Y. J. T. S. Xue, and Technology, "Droiddetector: android malware characterization and detection using deep learning," vol. 21, no. 1, pp. 114-123, 2016.

[78] Y. Li, R. Ma, R. J. I. J. o. S. Jiao, and I. Applications, "A hybrid malicious code detection method based on deep learning," vol. 9, no. 5, pp. 205-216, 2015.

[79] R. Vinayakumar, M. Alazab, K. Soman, P. Poornachandran, and S. Venkatraman, "Robust intelligent malware detection using deep learning," *IEEE Access,* vol. 7, pp. 46717-46738, 2019.

[80] A. Damodaran, F. Di Troia, C. A. Visaggio, T. H. Austin, and M. Stamp, "A comparison of static, dynamic, and hybrid analysis for malware detection," *Journal of Computer Virology and Hacking Techniques,* vol. 13, no. 1, pp. 1-12, 2017.