A Survey On Machine Learning For Cyber Security

A. Lakshmanarao, M. Shashi

Abstract: Cyber crime is proliferating everywhere exploiting every kind of vulnerability to computing environment. Ethical Hackers pay more attention towards assessing vulnerabilities and recommending mitigation methodologies. The development of effective techniques has been an urgent demand in the field of the cybersecurity community. Machine Learning for cybersecurity has become an issue of great importance recently due to the effectiveness of machine learning and deep learning in cybersecurity issues. Machine learning techniques have been applied for major challenges in cybersecurity issues like intrusion detection, malware classification and detection, spam detection and phishing detection. Although machine learning cannot automate a complete cybersecurity system, it helps to identify cyber-security threats more efficiently than other software-oriented methodologies, and thus reduces the burden on security analysts. Ever evolving nature of cyber threats throws challenges continuously on the researchers to explore with the ideal combination of deep expertise in cybersecurity and in data science. In this paper, we present the current state of art machine learning applications and their potential for cybersecurity. An analysis of machine learning algorithms for most common types of cybersecurity threats is presented.

Index Terms: Cybersecurity, Malware detection, Machine learning, Deep learning.

1 INTRODUCTION

Since the invention of the internet technology, cyberspace has emerged as a cen-tral hub for the creation of cyberattacks. The advances in technologies further facilitate hackers to discover vulnerabilities and to create viruses and malware continuously challenging the cyber security industry. Cyber security involves providing secure computing and communicative environment with proper innovations and procedures intended to shield PCs, systems, projects, and information from assault, unapproved access, change, or annihilation. These frameworks are made out of network security and host security systems with firewalls, anti-virus softwares, Intrusion detection systems etc. Machine Learning is proven to be capable of solving the most common problems in different domains like image processing, Health informatics applications, physical sciences, Computational Biology, Robotics, Financial prediction, Audio Processing, Medical Diagnostics, Video Processing, Text Processing [1].Specifically Machine learning techniques are also applied successfully in the field of cybersecurity to develop effective solutions. Machine learning has excellent potential for detecting various types of cyber-attacks and thus has become an important tool for the defenders. ESET conducted a survey on “usage of machine learning for cybersecurity”, in which 80% of the participants believed that Machine Learning will help their organization to detect and respond faster to threats [9].

2 MACHINE LEARNING TECHNIQUES:

2.1 Regression:

In regression, the value of a dependent feature is estimated based on the values of the independent features by learning from the existing data related to past events and such knowledge is used to handle new events. In cybersecurity, Fraud detection can be solved by regression. Once a model is learnt from the past transaction database, based on observed features of the current transactions, it determines fraudulent transactions. Machine learning provides Linear regression, Polynomial regression, Support vector machine, Decision tree, Random forest and other regression methods for regression analysis. Venkatesh Jaganathan [2] et.al applied multiple regression techniques for predicting the impact of attacks. They have taken the Overall CVSS (Common Vulnerability Scoring System) score as a dependent variable and two independent variables as X1(number of vulnerabilities), X2(Average Input Network Traffic). Daria Lavrova [3] et.al proposed a multiple regression model for the detection of security incidents in the IoT. With this technique, they were able to find known and unknown attacks.

2.2 Classification:

Classification is another extensively used supervisory machine learning task. In cybersecurity, spam detection is successfully implemented by ML based classifiers which involves discriminating a given email messages as spam or not. The spam filter models are able to separate spam messages from non-spam messages. Machine learning techniques for classification include Logistic Regression,K-Nearest Neighbors, Support Vector Machine, Naïve Bayes,Decision Tree,Random Forest Classification.Upon the availability of large collection of past data with labels, Deep Learning classification models involving Restricted Boltzmann Machines(RBM), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), or Long-Short Term Memory (LSTMs) cells for feature extraction followed by a densely connected neural network have become more efficient in solving complex tasks. Applicability of the above supervisory machine learning techniques is conditioned based on the availability of large collection of labeled data.

2.3 Clustering:

Both regression and classification are supervised learning models, for which labeled data is essential. Clustering is an

- *A. Lakshmanarao, Assistant Professor, Department of CSE, Raghu Engineering College, Dakamarri, Visakhapatnam, A.P,India, Email: laxman1216@gmail.com*
- *M.Shashi, Professor, Department of CS & SE, College of Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India, Email: smogalla2000@yahoo.com*

unsupervised learning model, which extracts general patterns from the data even when the data is not labeled. Groups of similar events constitute a cluster as they share common features that define a specific(behavior)pattern. In cybersecurity, clustering can be used for forensic analysis, anomaly detection, malware analysis, etc. K-means, K-Medoids, DBSCAN, Gaussian Mixture Model, Agglomerative clustering are some of the ML clustering techniques used in cybersecurity. Neural network based Self Organizing Maps (SOMs) can also be used for clustering.

3 CYBER SECURITY ISSUES:

The four major areas where Machine Learning algorithms play a crucial role are Intrusion Detection Systems, Malware analysis, Mobile (Android) malware detection and Spam Detection.

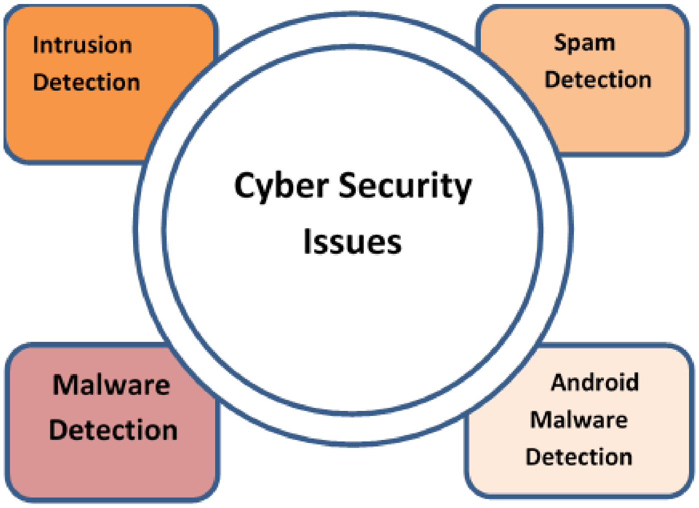

***Figure 1:** Cyber security issues*

3.1 Intrusion Detection:

Whenever secure information compromised by malicious software or policy violations, then Intrusion Detection Systems comes into the picture. Detection of an intrusion can be done in several ways. Broadly the methods are classified into either signature-based or anomaly-based. In the signature-based approach, all packets are compared with the signatures of known malicious threats. In the anomaly-based approach, network traffic is monitored against an established baseline of normality. Saroj Kr. Biswas [4] showed that machine learning based feature selection techniques play an important role in a good intrusion detection system. They applied a combination of feature selection techniques and achieved good results. R. Vinaya Kumar [5] et.al proposed a scale-hybrid-IDS-AlertNet system that can analyze network, host-level activities. This model was created using Deep Neural Networks (DNNs). They developed a scalable framework which is based on big data approaches and Apache Spark cluster computing platform. They conducted different experiments using DNNs with 1000 epochs and learning rate between 0.01 to 0.5 on various publicly available datasets like CICIDS 2017, KDDCup 99, UNSW-NB15, NSL-KDD, WSN-DS. They also applied traditional machine learning algorithms as baselines for comparisons. Md.Zahangir Alom [6] proposed a Deep Belief Networks for Intrusion Detection and compared their model with the SVM. The features are extracted from the training set using a two-layer Restricted Boltzmann Machine (RBM). The deep belief networks-based IDS could outperform SVM model and achieved an accuracy of 97.5%J. Kim [7] et.al applied a specific type of Recurrent Neural Networks called LSTM model for training the IDS using KDD Cup 1999 dataset. They studied the impact of learning rate and the number of neurons in the hidden layer upon the attack Detection Rate. They conducted several experiments with different learning rate and hidden layer sizes and achieved Detection Rate of 98.88%. Anna L. Buczak et.al [8] stated that data (pcap, NetFlow, or other network data) play a vital role in applying ML/DM approach for Intrusion Detection System. They also noted that there is a large gap in the availability of labeled data. N. Shone [10] et.al proposed a deep learning model for Network Intrusion Detection System operation with combination of ML and DL methods. They proposed a non-symmetric deep autoencoder (NDAE) on KDD Cup ’99, NSL-KDD datasets.

3.2 Malware Detection:

Malware is coined from ‘Malicious software’ in short,is a specific type of cyber threat software.Generally it is used for illegal activities like compromising the system by stealing data or bypass access control or cause harm to the host computer and the like.The term malware is broadly used for various types of malicious programs like viruses, Trojan horses, worms, bugs, adware, bots, rootkits, spyware, Ransomware, Key logger, backdoor. Each of these malware types contain several families. For example, ransomware can be classified as Charger family, Jisut family, Koler family, Pletor family, RansomBO family, Svpeng family, Simplocker family, etc. These malicious programs can be embedded in different formats like UNIX ELF (Executable and linkable) files, windows PE files (Portable Executables with .exe, dll, efi.). Document-based malware programs can be embedded in .doc,.pdf,.rtf files. Malware can also be in the form of extensions and plugins for popular software platforms like web browsers, web frameworks. Dolly Uppal [11] et.al proposed a malware classification and detection system based on the n-gram method. They have applied a pre-modeled program for tracking the execution of the samples and captured the API calls. After generating the feature vector, they applied different machine learning algorithms and achieved the best results with the SVM classifier.Mozammel Chowdhury [12] et.al proposed a Neural Network-based approach for malware detection. They extracted features from PE headers using the n-gram method and conducted experiments with the extended set of features and achieved 97% accuracy with ANN. Bowen Sun [13] et.al proposed a malware classification model using static features in different perspectives. They extracted static features in 3 perspectives including PE features, bytecode features, and assembler code features. They compared the performance of eight classifiers among which the best classifier could achieve an f1-score of 93.56%. Mahmoud Kalash [14] et.al proposed a CNN for malware classification. They represented the codes of 25 families of malware binaries to grayscale images and applied CNN for classification. They conduct experiments with two well-known datasets ‘Malimg’ and ‘Microsoft malware’ and reported that they achieved 98.52%, 99.97% accuracy on the two datasets respectively.

3.3 Android Malware Detection:

Android is the most widely used mobile platform and hence highly targeted by the mobile malware creators. As the number of android malware types are increasing day by day, it has become more and more challenging to detect and classify mobile malware variants. A large number of attempts are made by the researchers towards mobile malware detection. DroidMat [15] applied k-means clustering and K-NN algorithms on static features from android apps. Arp et al. [16], Varsha et al. [17], Sharma and Dash [18] extracted static features from android apps and they achieved good results by applying machine algorithms like SVM, Random Forest, K-NN, Naive Bayes, Decision Trees. AntiMalDroid [19],Droid Dolphin [20] applied Support Vector Machines on dynamic features extracted from malware apps (logged behavior sequence as features) and achieved good accuracy. Suleiman Y. Yerima [21] et al. proposed a Multilevel Classifier Fusion method for Android Malware Detection. They proposed four ranking based algorithms based on accuracy, recall and precision rates. Based on their ranking algorithms, they combined four classifiers to achieve a better detection rate. They evaluated their model performance on three datasets and achieved a good recall rate.

3.4 Spam Detection:

Spam Detection is also one of the major challenges in cybersecurity. Spam is an unsolicited bulk messaging generally used for advertising. Generally, spam indicates email spam, but it could be a message on social networking sites and other blogging platforms also. Spam messages waste a lot of valuable time. Sometimes, users get spam emails that disguised themselves as authentic message from a bank to trap the users. Responding to such spam messages may lead to incur heavy financial lossess. Machine learning techniques have been applied by many researchers for spam detection. Muhammad N. Marsono[22] et al applied the Naïve Bayes classification technique for identification of spam messages among incoming email and achieved good results. James Clark [23] et al applied the K-NN model for automated email classification problem. S. Jancy Sickory Daisy[24] proposed a hybrid spam detection system based on Naive Bayes classification and Markov Random Field method. They evaluated their model based on its accurateness, time consumption and claimed that the performance of the hybrid approach is better than the baseline methods. Sreekanth Madisetty [25] et.al proposed an ensemble model for spam classification on Twitter. They developed deep learning models based on CNNs.They ap-plied various word embedding to preprocess the input in textual form into numeric form before training the CNN model.They used 5 CNNs (CNN + Twitter Glove, CNN + Google News, CNN + Edinburgh, CNN + H Spam, CNN + Random) for word embeddings and one feature-based model for spam detection. Mehul Gupta [26] et.al compared various machine learning and deep learning techniques for SMS spam detection on two different datasets. They compared the performance of eight different classifiers and showed that CNN Classifier achieved the accuracy of 99.19% and 98.25% for the two datasets. Figure 2 shows a summary of machine learning algorithms for solving various cybersecurity issues. Although most of the researchers applied all the machine learning algorithms for all four cybersecurity issues, we summarized only appropriate models for specific cybersecurity issue. Intrusion detection can be solved by good feature selection techniques and deep learning models like Recurrent Neural Networks (RNNs). Malware detection (PC) can be solved by ANNs and CNNs effectively. Malware samples are first converted to images and then CNN's are applied. Android malware detection can be addressed by shallow machine learning algorithms and various fusion models. Spam detection can be efficiently addressed by shallow machine learning models like Naïve Bayes and K-NN models and deep learning models like CNN.

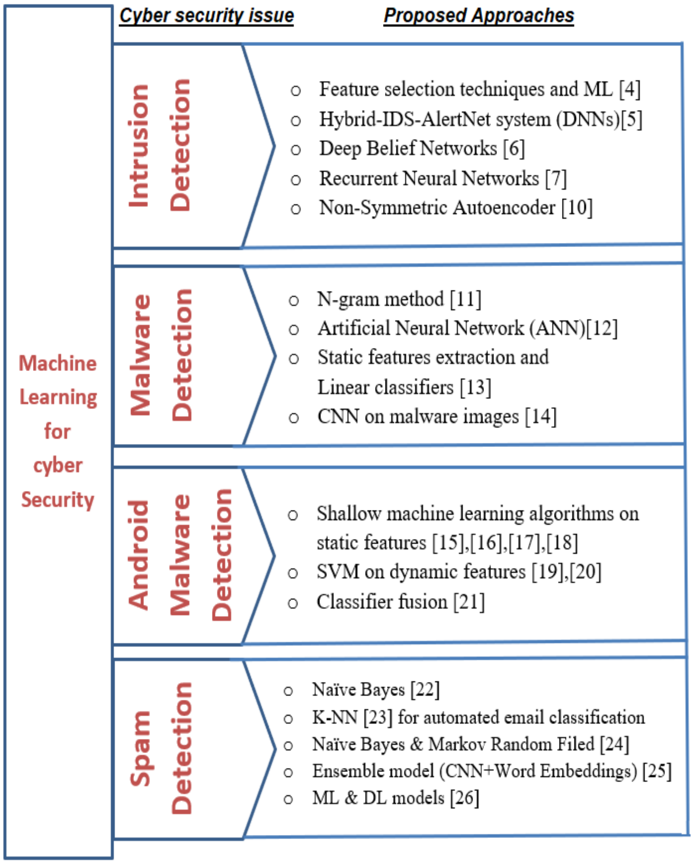
Cyber security issue
Proposed Approaches
Machine Learning for cyber Security
Intrusion Detection
○ Feature selection techniques and ML [4]
○ Hybrid-IDS-AlertNet system (DNNs)[5]
○ Deep Belief Networks [6]
○ Recurrent Neural Networks [7]
○ Non-Symmetric Autoencoder [10]
Malware Detection
○ N-gram method [11]
○ Artificial Neural Network (ANN)[12]
○ Static features extraction and Linear classifiers [13]
○ CNN on malware images [14]
Android Malware Detection
○ Shallow machine learning algorithms on static features [15],[16],[17],[18]
○ SVM on dynamic features [19],[20]
○ Classifier fusion [21]
Spam Detection
○ Naïve Bayes [22]
○ K-NN [23] for automated email classification
○ Naïve Bayes & Markov Random Filed [24]
○ Ensemble model (CNN+Word Embeddings) [25]
○ ML & DL models [26]

Figure 2: *Machine Learning for cyber security*

4 CONCLUSION

Machine Learning approaches are widely applied to solve various types of cybersecurity problems. Advances in the field of machine learning and deep learning offers promising solutions to cybersecurity issues. But it is important to identify which algorithm is suitable for which application. Multi-Layered approaches are needed to keep the solution resilient against malware attacks and to achieve high detection rates. The selection of a particular model plays a vital role in solving cybersecurity issues. In this paper the authors explored the state of art mechanisms for cybersecurity problems. The autonomous capabilities of machine learning and deep learning algorithms must not be overestimated. The combination of human supervision and Machine learning techniques results in achieving the desired goals of cybersecurity.

REFERENCES

[1] William G Hatcher, Wei Yu, “A Survey of Deep Learning: Platforms, Applications and Emerging Research Trends”,

IEEE Access 2018, Volume: 6, DOI:10.1109/ACCESS.2018.2830661.

[2] Venkatesh Jaganathan, Premapriya Muthu Sivashanmugam, Priyesh Cherurveettil, “Using a Prediction Model to Manage Cyber Security Threats”, Hindawi Publishing Corporation the Scientific World Journal Volume 2015, Article ID 703713, http://dx.doi.org/10.1155/2015/703713.

[3] Daria Lavrova, Alexander Pechenkin,” Applying Correlation and Regression Analysis to Detect Security Incidents in the Internet of Things”, International Journal of Communication Networks and Information Security (IJCNIS), Volume. 7, No. 3, December 2015.

[4] Saroj Kr. Biswas, “Intrusion Detection Using Machine Learning: A Comparison Study”, International Journal of Pure and Applied Mathematics, Volume 118 No. 19 2018, 101-114.

[5] R. Vinayakumar, Mamoun Alazab, (Senior Member, IEEE), K. P. Soman, Prabaharan Poornachandran, Ameer Al-Nemrat, A.N. Venkatraman, “Deep Learning Approach for Intelligent Intrusion Detection System”, IEEE Access, VOLUME 7, 2019, Digital Object Identifier 10.1109/ACCESS.2019.2895334.

[6] Md. Zahangir Alom, Venkata Ramesh Bontupalli, and Tarek M. Taha, “Intrusion Detection using Deep Belief Networks”, 978-1-4673-7565-8/15/$31.00 ©2015 IEEE

[7] J. Kim, L. T. Thu and H. Kim “Long Short-Term Memory Recurrent Neural Network Classifier for Intrusion Detection,” IEEE International Conference on Platform Technology and Service, 2016.

[8] Anna L. Buczak and Erhan Guven,” A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection”, IEEE Communications Surveys and Tutorials, Volume. 18, No. 2,2nd Quarter 2016.

[9] Ondrej Kubovič (ESET Security Awareness Specialist),” Machine-Learning Era in Cy-bersecurity: A Step Towards A Safer World or The Brink of Chaos”, Machine-Learning Era in Cybersecurity White Paper, February 2019

[10] N. Shone, V. D. Phai, T. N. Ngoc, Q. Shi, "A deep learning approach to network intrusion detection", IEEE Transactions on Emerging Topics in Computational Intelligence-Feb-2018(41-50).

[11] Dolly Uppal, Vinesh Jain, Rakhi Sinha and Vishakha Mehra and “Malware Detection and Classification Based on Extraction of API Sequences”, 978-1-4799-3080-7/14/$31.00_c 2014 IEEE.

[12] Mozammel Chowdhury, Azizur Rahman, Rafiqul Islam, “Protecting Data from Mal-ware Threats using Machine Learning Technique”, 2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA).

[13] Bowen Sun, Qi Li, Yanhui Guo, Qiaokun Wen, Xiaoxi Lin, Wenhan Liu, “Malware Family Classification Method Based on Static Feature Extraction”, 2017 3rd IEEE International Conference on Computer and Communications

[14] Mahmoud Kalash, Mrigank Rochan, Noman Mohammed, Neil D. B. Bruce, Yang Wang, Farkhund Iqbal, “Malware Classification with Deep Convolutional Neural Net-works”, 978-1-5386-3662-6/18/$31.00 ©2018 IEEE

[15] D.-J. Wu, C.-H. Mao, T.-E. Wei, H.-M. Lee, and K.-P. Wu, “DroidMat: Android mal-ware detection through manifest and API calls tracing,” in Proc. 7th Asia Joint Conf. Inf. Security (Asia JCIS), 2012, pp. 62–69.

[16] D. Arp, M. Spreitzenbarth, M. Hubner, H. Gascon, and K. Rieck, “Drebin: Efficient and explainable detection of Android malware in your pocket,” in Proc. 20th Annu. Netw. Distrib. Syst. Security Symp. (NDSS), San Diego, CA, USA, Feb. 2014, pp. 1–15.

[17] M. V. Varsha, P. Vinod, and K. A. Dhanya, “Identification of malicious Android app using manifest and opcode features,” J. Comput. Virol. Hacking Tech., vol. 13, no. 2, pp. 125–138, 2017.

[18] A. Sharma and S. K. Dash, “Mining API calls and permissions for Android malware detection,” in Cryptology and Network Security. Cham, Switzerland: Springer Int., 2014, pp. 191–205.

[19] M. Zhao, F. Ge, T. Zhang, and Z. Yuan.,” An efficient SVM-based malware detection framework for Android,” in Communications in Computer and Information Science, vol. 243, Springer, 2011, pp. 158–166.

[20] W.-C. Wu, S.-H. Hung, “A dynamic Android malware detection framework using big data and machine learning,” in Proc. ACM Conf. Res. Adapt. Convergent Syst. (RACS), Towson, MD, USA, 2014, pp. 247–252.

[21] Suleiman Y. Yerima, Member, IEEE, and Sakir Sezer, Member, IEEE, “Droid Fusion: A Novel Multilevel Classifier Fusion Approach for Android Malware Detection”, IEEE TRANSACTIONS ON CYBERNETICS, VOL. 49, NO. 2, FEBRUARY 2019.

[22] Muhammad N. Marsono, M. Watheq El-Kharashi, Fayez Gebali, “Targeting spam control on middleboxes: Spam detection based on layer-3 e-mail content classification” Elsevier Computer Networks, 2009.

[23] James Clark, Irena Koprinska, Josiah Poon, “A Neural Network Based Approach to Automated E-mail Classification”, Proceedings IEEE/WIC International Conference on Web Intelligence, 0-7695-1932-6, Oct. 2003.

[24] S. Jancy Sickory Daisy, A.Rijuvana Begum, “Hybrid Spam Filtration Method using Ma-chine Learning Techniques”, International Journal of Innovative Technology and Exploring Engineering, ISSN: 2278-3075, Volume-8, Issue-9, July 2019.

[25] Sreekanth Madisetty and Maunendra Sankar Desarkar, “A Neural Network-Based Ensemble Approach for Spam Detection in Twitter”, IEEE Transactions on Computational Social Systems, Volume: 5, Issue: 4, Dec. 2018.

[26] Mehul Gupta, Aditya Bakliwal, Shubhangi Agarwal & Pulkit Mehndiratta, “A Comparative Study of Spam SMS Detection using Machine Learning Classifiers”, Eleventh International Conference on Contemporary Computing (IC3), 2-4 August, 2018, Noida, India, 978-1-5386-6835-1/18,2018 IEEE