Effectiveness of Machine & Deep Learning For Cyber security

Shamal Tavandkar[1]
Department of Computer Engineering
[1] Siddhant College Of Engineering, Sudumbre, Pune

Abstract- *With the development of the Internet, cyber-attacks are changing rapidly and the cyber security situation is not optimistic. This survey report describes key literature surveys on machine learning (ML) and deep learning (DL) methods for network analysis of intrusion detection and provides a brief tutorial description of each ML / DL method. Papers representing each method were indexed, read, and summarized based on their temporal or thermal correlations. Because data are so important in ML / DL methods, we describe some of the commonly used network datasets used in ML / DL, discuss the challenges of using ML / DL for cyber security and provide suggestions for research directions*

Keywords- Cyber security, Intrusion detection, Deep learning, Machine learning

I. INTRODUCTION

With the increasingly in-depth integration of the Internet and social life, the Internet is changing how people learn and work, but it also exposes us to increasingly serious security threats. How to identify various network attacks, particularly not previously seen attacks, is a key issue to be solved urgently.

Cyber security is a set of technologies and processes designed to protect computers, networks, programs and data from attacks and unauthorized access, alteration, or destruction. A network security system consists of a network security system and a computer security system. Each of these systems includes firewalls, antivirus software, and intrusion detection systems (IDS). IDSs help discover, determine and identify unauthorized system behavior such as use, copying, modification and destruction.

Security breaches include external intrusions and internal intrusions. There are three main types of network analysis for IDSs: misuse-based, also known as signature-based, anomaly-based, and hybrid. Misuse-based detection techniques aim to detect known attacks by using the signatures of these attacks. They are used for known types of attacks without generating a large number of false alarms. However, administrators often must manually update the database rules and signatures. New (zero-day) attacks cannot be detected based on misused technologies.

Anomaly-based techniques study the normal network and system behavior and identify anomalies as deviations from normal behavior. They are appealing because of their capacity to detect zero-day attacks. Another advantage is that the profiles of normal activity are customized for every system, application, or network, therefore making it difficult for attackers to know which activities they can perform undetected. Additionally, the data on which anomaly-based techniques alert (novel attacks) can be used to define the signatures for misuse detectors. The main disadvantage of anomaly-based techniques is the potential for high false alarm rates because previously unseen system behaviors can be categorized as anomalies.

Hybrid detection combines misuse and anomaly detection[4]. It is used to increase the detection rate of known intrusions and to reduce the false positive rate of unknown attacks. Most ML / DL methods are hybrids.

This paper presents a literature review of machine learning (ML) and deep learning (DL) methods for cybersecurity applications. ML / DL methods and some applications of each method in network intrusion detection are described. It focuses on ML and DL technologies for network security, ML/DL methods and their descriptions. Our research aims on standards-compliant publications that use "machine learning", "deep learning" and cyber as keywords to search on Google Scholar. In particular, the new hot papers are used because they describe the popular techniques.

The purpose of this paper is for those who want to study network intrusion detection in ML/DL.Thus, great emphasis is placed on a thorough description of the ML/DL methods, and references to seminal works for each ML and DL method are provided. Examples are provided concerning how the techniques were used in cyber security.

This paper does not describe all of the different techniques of network anomaly detection; instead, it concentrates only on ML and DL techniques. However, in addition to anomaly detection, signature-based and hybrid methods are depicted.

Patcha et al. [5] discuss technological trends in anomaly detection and identify open problems and challenges in anomaly detection systems and hybrid intrusion detection systems. However, their survey only covers papers published from 2002 to 2006, whereas our survey includes more-recent papers. Unlike Modi C et al. [6], this review covers the application of ML / DL in various areas of intrusion detection and is not limited to cloud security.

Revathi S et al.[7] focus on machine-learning intrusion techniques. The authors present a comprehensive set of machine-learning algorithms on the NSL-KDD intrusion detection dataset, but their study only involves a misuse detection context. In contrast, this paper describes not only misuse detection but also anomaly detection.

II. SIMILARITIES AND DIFFERENCES IN ML AND DL

There are many puzzles about the relationship among ML, DL, and artificial intelligence (AI). AI is a new technological science that studies and develops theories, methods, techniques, and applications that simulate, expand and extend human intelligence. It is a branch of computer science that seeks to understand the essence of intelligence and to produce a new type of intelligent machine that responds in a manner similar to human intelligence. Research in this area includes robotics, computer vision, nature language processing and expert systems. AI can simulate the information process of human consciousness, thinking. AI is not human intelligence, but thinking like a human might also exceed human intelligence.

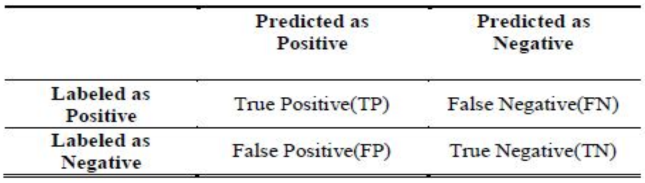
ML is a branch of AI and is closely related to (and often overlaps with) computational statistics, which also focuses on prediction making using computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. ML is occasionally conflated with data mining, but the latter subfield focuses more on exploratory data analysis and is known as unsupervised learning. ML can also be unsupervised and be used to learn and establish baseline behavioral profiles for various entities and then used to find meaningful anomalies. The pioneer of ML, Arthur Samuel, defined ML as a “field of study that gives computers the ability to learn without being explicitly programmed.” ML primarily focuses on classification and regression based on known features previously learned from the training data.

DL is a new field in machine-learning research. Its motivation lies in the establishment of a neural network that simulates the human brain for analytical learning. It mimics the human brain mechanism to interpret data such as images, sounds and texts.

The concept of DL was proposed by Hinton et al. based on the deep belief network (DBN), in which an unsupervised greedy layer-by-layer training algorithm is proposed that provides hope for solving the optimization problem of deep structure. Then the deep structure of a multi-layer automatic encoder is proposed. In addition, the convolution neural network proposed by Lecun et al is the first real multi-layer structure learning algorithm that uses a space relative relationship to reduce the number of parameters to improve the training performance.

DL is a machine-learning method based on characterization of data learning. An observation, such as an image, can be expressed in a variety of ways, such as a vector of each pixel intensity value, or more abstractly as a series of edges, a region of a particular shape, or the like. Using specific representations makes it easier to learn tasks from instances. Similarly to ML methods, DL methods also have supervised learning and unsupervised learning. Learning models built under different learning frameworks are quite different. The benefit of DL is the use of unsupervised or semi-supervised feature learning and hierarchical feature extraction to efficiently replace features manually.

TABLE 1. Confusion matrix.



	Predicted as Positive	Predicted as Negative
Labeled as Positive	True Positive(TP)	False Negative(FN)
Labeled as Negative	False Positive(FP)	True Negative(TN)

For a binary classification as shown in Table 1, the results can be divided into four categories:

- True Positive (TP): Positive samples correctly classified by the model;
- False Negative (FN): A positive sample that is mis-classified by the model;
- False Positive (FP): A negative samples that is mis-classified by the model;
- True Negative (TN): Negative samples correctly classified by the model;
- Further, the following metrics can be calculated from the confusion matrix:
- Accuracy: (TP + TN)/ (TP + TN + FP + FN). Ratio of the number of correctly classified samples to the total number

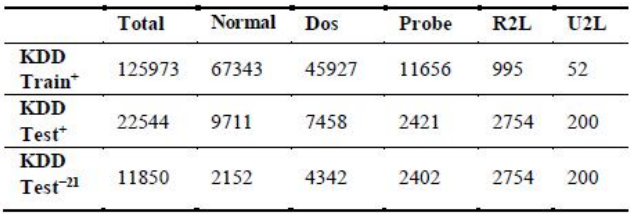
of samples for a given test data set. When classes are balanced, this is a good measure; if not, this metric is not very useful.

- Precision: TP/ (TP + FP). It calculates the ratio of all "correctly detected items" to all "actually detected items".
- Sensitivity or Recall or True Positive Rate (TPR): TP/ (TP + FN). It calculates the ratio of all "correctly detected items" to all "items that should be detected".
- False Negative Rate (FNR): FN/ (TP + FN). The ratio of the number of misclassified positive samples to the number of positive samples.
- False Positive Rate (FPR): FP/ (FP + TN). The ratio of the number of misclassified negative samples to the total number of negative samples.
- True Negative Rate (TNR): TN/ (TN + FN). The ratio of the number of correctly classified negative samples to the number of negative samples.
- F1-score: 2*TP/(2*TP + FN + FP). It calculates the harmonic mean of the precision and the recall.

III. NETWORK SECURITY DATA SET

Data constitute the basis of computer network security research. The correct choice and reasonable use of data are the prerequisites for conducting relevant security research. The size of the dataset also affects the training effects of the ML and DL models. Computer network security data can usually be obtained in two ways: 1) directly and 2) using an existing public dataset. Direct access is the use of various means of direct collection of the required cyber data, such as through Win Dump or Wire shark software tools to capture network packets. This approach is highly targeted and suitable for collecting short-term or small amounts of data, but for long-term or large amounts of data, acquisition time and storage costs will escalate. The use of existing network security datasets can save data collection time and increase the efficiency of research by quickly obtaining the various data required for research. This section will introduce some of the Security datasets that are accessible on the Internet and facilitate section IV of the research results based on a more comprehensive understanding.

TABLE 2. Different classifications in the NSL-KDD



	Total	Normal	Dos	Probe	R2L	U2L
KDD Train^{+}	125973	67343	45927	11656	995	52
KDD Test^{+}	22544	9711	7458	2421	2754	200
KDD Test^{-21}	11850	2152	4342	2402	2754	200

The dataset covers the KDD Train + dataset as the training set and KDD Test + and KDD Test − 21 datasets as the testing set, which has different normal records and four different types of attack records, as shown in Table 2. The KDDTest−21 dataset is a subset of the KDD Test+ and is more difficult to classify.

IV. ML and DL Algorithm for Cyber Security

This section is divided into two parts. The first part introduces the application of traditional machine-learning algorithms in network security. The second part introduces the application of deep learning in the field of cyber security. It not only describes the research results but also compares similar studies.

A. SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is one of the most robust and accurate methods in all machine-learning algorithms. It primarily includes Support Vector Classification (SVC) and Support Vector Regression (SVR). The SVC is based on the concept of decision boundaries. A decision boundary separates a set of instances having different class values between two groups. The SVC supports both binary and multi-class classifications. The support vector is the closest point to the separation hyperplane, which determines the optimal separation hyperplane. In the classification process, the mapping input vectors located on the separation hyperplane side of the feature space fall into one class, and the positions fall into the other class on the other side of the plane. In the case of data points that are not linearly separable, the SVM uses appropriate kernel functions to map them into higher dimensional spaces so that they become separable in those spaces

B. K-NEARESTNEIGHBOR

The kNN classifier is based on a distance function that measures the difference or similarity between two instances. The standard Euclidean distance d(x, y) between two instances x and y is defined as :

n 2 k k k=1 d(x,y)= (x -y) where, xk is the kth featured element of instance x, yk is the kth featured element of the instance y and n is the total number of features in the dataset.

Assume that the design set for kNN classifier is U. The total number of samples in the design set is S. Let C = {C1 ,C2 ,…CL} are the L distinct class labels that are available in S. Let x be an input vector for which the class label must be predicted. Let yk denote the kth vector in the design set S. The kNN algorithm is to find the k closest vectors in design set S to input vector x. Then the input vector x is classified to class

Cj if the majority of the k closest vectors have their class as Cj.

Rao used Indexed Partial Distance Search k- Nearest Neighbor (IKPDS) to experiment with various attack types and different k values (i.e., 3, 5, and 10). They randomly selected 12,597 samples from the NSl-KDD dataset to test the classification results, resulting in 99.6% accuracy and faster classification time. Experimental results show that IKPDS, and in a short time Network Intrusion Detection Systrms(NIDS), have better classification results. However, the study of the test indicators of the experiment is not perfect; it did not consider the precision and recall rate.

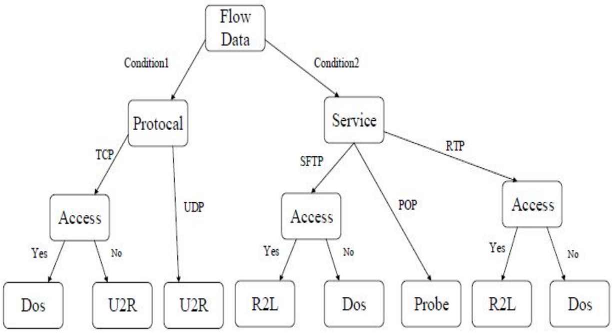


FIGURE 1. An example decision tree

C. DECISION TREE

A decision tree is a tree structure in which each internal node represents a test on one property and each branch represents a test output, with each leaf node representing a category. In machine learning, the decision tree is a predictive model; it represents a mapping between object attributes and object values. Each node in the tree represents an object, each divergence path represents a possible attribute value, and each leaf node corresponds to the value of the object represented by the path from the root node to the leaf node. The decision tree only has a single output; if you want complex output, you can establish an independent decision tree to handle different outputs. Commonly used decision tree models are ID3, C4.5 and CART.

As shown in Fig.1, the decision tree classifies the samples through the conditions of training, and has better detection accuracy for known intrusion methods, but is not suitable for detection of unknown intrusion.

Ingre and Bhupendra propose a decision tree-based IDS for the NSL-KDD dataset. Feature selection using a correlation feature selection (CFS) approach, selecting 14 methods were used for experiments. The experiments were performed on a KDD99 dataset.

D. DEEP BELIEF NETWORK

Deep Belief Network (DBN) is a probabilistic generative model consisting of multiple layers of stochastic and hidden variables. The Restricted Boltzmann Machine (RBM) and DBN are interrelated because composing and stacking a number of RBMs enables many hidden layers to train data efficiently through activations of one RBM for further training stages [56]. RBM is a special topological structure of a Boltzmann machine (BM). The principle of BM originated from statistical physics as a modeling method based on an energy function that can describe the high-order interactions between variables. BM is a symmetric coupled random feedback binary unit neural network composed of a visible layer and a plurality of hidden layers. The network node is divided into a visible unit and a hidden unit, and the visible unit and the hidden unit are used to express a random network and a random environment. The learning model expresses the correlation between units by weighting.

E. RECURRENT NEURAL NETWORKS

The recursive neural network (RNN) is used to process sequence data. In the traditional neural network model, data from the input layer to the hidden layer to the output layer; The layers are fully connected and there is no connection between the nodes between each layer. Many problems exist that this conventional neural network cannot solve.

F. COVOLUTIONAL NEURAL NETWORKS

The recursive neural network (RNN) is used to process sequence data. In the traditional neural network model, data from the input layer to the hidden layer to the output layer; The layers are fully connected and there is no connection between the nodes between each layer. Many problems exist that this conventional neural network cannot solve.

Convolutional Neural Networks (CNN) is a type of artificial neural network that has become a hotspot in the field of speech analysis and image recognition. Its weight-sharing network structure makes it more similar to a biological neural network, thus reducing the complexity of the network model and reducing the number of weights. This advantage is more obvious when the network input is a multi-dimensional image, and the image can be directly used as the input of the network to avoid the complex feature extraction and data reconstruction in the traditional recognition algorithm. The Convolutional Network is a multi-layered sensor specifically designed to recognize two-dimensional shapes that are highly invariant to translation, scaling, tilting, or other forms of deformation.

CNN is the first truly successful learning algorithm for training multi-layer network structures, that is, the structure shown in Fig 3. It reduces the number of parameters that must be learned to improve the training performance of the BP algorithm through spatial relationships. As a deep learning

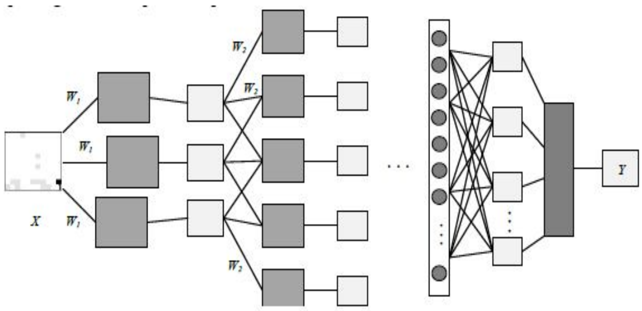


FIGURE 3. An example CNN model structure.

V. DISCUSSION AND FUTURE DIRECTION

A. DATA SETS

Existing datasets have the defects of old data, redundant information and unbalanced numbers of categories. Although the data can be improved after processing, there is a problem of insufficient data volume. Therefore, establishing network intrusion detection datasets with large amounts of data, wide-type coverage and balanced sample numbers of attack categories become a top priority in the field of intrusion detection.

B. HYBRID METHOD

Hybrid detection methods mostly combine machine-learning methods such as those described by [30,33,40], whereas intrusion detection with a combination of deep learning and machine-learning methods is less studied. AlphaGo has validated the validity of this idea, which is an exciting research direction.

C. DETECTION SPEED

By reducing the detection time and improving the detection speed from the algorithm and hardware aspects, the algorithm can be used less time given the complexity of the machine-learning algorithm and deep learning algorithm. Hardware can use multiple computers for parallel computing. Combining the two approaches is also an interesting topic.

D. ONLINE LEARNING

The means of network intrusion is increasing day by day. How to fit the new data better with the trained model is also an exciting research direction. At present, transfer learning is a viable means to fine-tune the model with a small amount of labeled data, which should be able to achieve better results in actual network detection.

VI. CONCLUSION

This paper presents a literature review of ML and DL methods for network security. The paper, which has mostly focused on the last three years, introduces the latest applications of ML and DL in the field of intrusion detection. Unfortunately, the most effective method of intrusion detection has not yet been established. Each approach to implementing an intrusion detection system has its own advantages and disadvantages, a point apparent from the discussion of comparisons among the various methods. Thus, it is difficult to choose a particular method to implement an intrusion detection system over the others.

Datasets for network intrusion detection are very important for training and testing systems. The ML and DL methods do not work without representative data, and obtaining such a dataset is difficult and time-consuming. However, there are many problems with the existing public dataset, such as uneven data, outdated content and the like. These problems have largely limited the development of research in this area.

Network information update very fast, which brings to the DL and ML model training and use with difficulty, model needs to be retrained long-term and quickly. So incremental learning and lifelong learning will be the focus in the study of this field in the future.

REFERENCES

[1] S. Aftergood, “Cybersecurity: The cold war online,” Nature, vol. 547, no. 7661, p. 30, 2017.

[2] A. Milenkoski, M. Vieira, S. Kounev, A. Avritzer, and B. D. Payne, “Evaluating Computer Intrusion Detection Systems:A Survey of Common Practices,” Acm Comput. Surv., vol. 48, no. 1, pp. 1–41, 2015.

[3] C. N. Modi and K. Acha, “Virtualization layer security challenges and intrusion detection/prevention systems in cloud computing: a comprehensive review,” J. Supercomput., vol. 73, no. 3, pp. 1–43, 2016.

[4] E. Viegas, A. O. Santin, A. França, R. Jasinski, V. A. Pedroni, and L. S. Oliveira, “Towards an Energy-Efficient Anomaly-Based Intrusion Detection Engine for Embedded Systems,” IEEE Trans. Comput., vol. 66, no. 1, pp. 163–177, 2017.

[5] A. Patcha and J. M. Park, “An overview of anomaly detection techniques: Existing solutions and latest technological trends,” Comput. Netw., vol. 51, no. 12, pp. 3448–3470, 2007.

[6] C. Modi, D. Patel, B. Borisaniya, H. Patel, A. Patel, and M. Rajarajan, “Review: A survey of intrusion detection techniques in Cloud,” J. Netw. Comput. Appl., vol. 36, no. 1, pp. 42–57, 2013.

[7] S. Revathi and A. Malathi, “A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection,” in International Journal of Engineering Research and Technology, 2013.

[8] D. Sahoo, C. Liu, and S. C. H. Hoi, “Malicious URL Detection using Machine Learning: A Survey,” arXiv:1701.07179, 2017.

[9] A. L. Buczak and E. Guven, “A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection,” IEEE Commun. Surv. Tutor., vol. 18, no. 2, pp. 1153–1176, 2016.

[10] M. Soni, M. Ahirwa, and S. Agrawal, “A Survey on Intrusion Detection Techniques in MANET,” in International Conference on Computational Intelligence and Communication Networks, 2016, pp. 1027–1032.