International Journal of Advanced Research in Engineering Technology & Science
ISSN: 2349-2819
Email: editor@ijarets.org Volume-8, Issue-10 October - 2021 www.ijarets.org

MACHINE LEARNING IN CYBER SECURITY

MOHAMMED NABEEL ZUBAIR

IIC University of Technology, Cambodia

Senior Specialist Network & Security Engineer

ABSTRACT

The progression of the Internet, the fast evolution of cyberattacks, and the pessimistic state of the current scenario in regards to cyber security. This paper contains a short instructional overview of each machine learning and deep learning approach, as well as major literature reviews on machine learning (ML) and deep learning (DL) methods for network analysis of intrusion detection. We indexed, reviewed, and summarised the papers that were representative of each approach based on the temporal or thermal correlations that they included. We outline some of the regularly used network datasets used in ML / DL, highlight the limitations of employing ML / DL for cybersecurity, and make ideas for future prospects since data play such a crucial role in ML / DL methodologies. Those individuals who are interested in doing research in the field of machine learning and deep learning are the target audience for this publication. As a consequence of this, a significant amount of attention is put on providing a comprehensive overview of the ML and DL procedures, and references to seminal publications for each of the ML and DL methods are supplied. The many applications of the approaches in the field of cyber security are shown with examples.This article does not provide a comprehensive overview of the many approaches to network anomaly detection; rather, it focuses only on machine learning and deep learning approaches. Nevertheless, in addition to detecting anomalies, signature-based techniques and hybrid approaches are presented here. Identify open difficulties and challenges in anomaly detection systems and hybrid intrusion detection systems, as well as explore the technical developments that are occurring in the field of anomaly detection. Their study, on the other hand, only covers publications that were published from 2002 to 2006, but our survey includes papers that were published more recently. This review, in contrast to Modi C et al .'s, examines the use of ML and DL in a variety of different aspects of intrusion detection and is not restricted to cloud security alone.

.***Keywords:*** *Cyber security, intrusion detection, deep learning, machine learning*

INTRODUCTION

The Internet is transforming the ways in which people learn and work, but it also puts us at risk of more grave dangers to our safety as it becomes ever more deeply integrated into all aspects of social life. An important problem that has to be addressed as soon as possible is how to recognise different types of network assaults, in particular ones that have not been observed before. Cybersecurity refers to a collection of technologies and methods that are intended to safeguard computers, networks, programmes, and data against intrusions as well as unwanted access, modification, or destruction. A computer security system and a network security system are the two components that make up a network security system. Every single one of these defences consists of a firewall, antivirus software, and intrusion detection systems (IDS). IDSs assist in the detection, determination, and identification of unlawful system activities

such as usage, copying, modification, and destruction.

Intrusions from the outside as well as those coming from inside are considered security breaches. There are three basic forms of network analysis for IDSs: misuse-based, also known as signature-based, anomaly- based, and hybrid. Misuse-based detection approaches try to identify known attacks by leveraging the fingerprints of these attacks .

SIMILARITIES AND DIFFERENCES IN ML AND DL

There are numerous unanswered questions about the connection between machine learning, deep learning, and artificial intelligence (AI). A new branch of computer science known as artificial intelligence (AI) focuses on the study and development of ideas, methodologies, techniques, and applications designed to imitate, extend, and expand human intellect. It is a subfield of computer science that aims to understand the fundamentals of intelligence and develop a new kind of intelligent machine that can react in a fashion that is analogous to the way that humans think intelligently. Robotics, computer vision, natural language processing, and expert systems are all examples of research that falls within this category. Artificial intelligence is capable of simulating the information processing of human mind and reasoning. Artificial intelligence is not the same as human intelligence, although thinking like a human may be more intelligent than humans.

ML is a subfield of AI that is closely connected to the field of computational statistics, which is likewise concerned with generating predictions with the use of computers and often overlaps with it. It has deep linkages to mathematical optimization, which supplies the discipline with methodologies, theoretical frameworks, and application fields. There are times when ML and data mining are confused with one another, however the latter area is known as unsupervised learning and relies more on exploratory data analysis than supervised learning does. Unsupervised machine learning may also be used to learn and construct baseline behavioural profiles for a variety of entities. This information can then be utilised to detect relevant abnormalities in the data. Arthur Samuel, a pioneer in the subject of machine learning, is credited with coining the term "machine learning" and defining it as "a branch of research that offers computers the capacity to learn without being explicitly taught." Machine learning focuses mostly on classification and regression, both of which are determined by using previously learned features from training data.

In the realm of machine learning research, DL is a relatively recent topic. The development of a neural network that can mimic the functions of the human brain in terms of analytical learning is the driving force behind it. It does this by imitating the technique that the human brain uses to analyse data such as pictures, sounds, and messages.

Hinton and colleagues came up with the idea of deep learning (DL) and based it on the deep belief network (DBN). Within the DBN, they proposed an unsupervised greedy layer-by-layer training algorithm that offers some glimmer of hope for finding a solution to the challenge of optimising deep structures. The subsequent step is to suggest the fundamental construction of a multi-layer automated encoder. In addition, Lecun et alconvolution .'s neural network is the first genuine multi-layer structure learning method that makes use of a space relative connection to cut down on the number of parameters in order to boost training performance. DL (or deep learning) is an approach to machine learning that is predicated on the characterisation of data learning. An observation, such as a picture, may be stated in a number of different

ways. For example, it might be expressed as a vector that contains the intensity value of each pixel; alternatively, it can be expressed more abstractly as a sequence of edges, an area of a certain form, or something similar. Learning tasks from their associated instances is simplified when certain representations are used. Both supervised and unsupervised learning are included in DL method workflows, just as they are in ML method workflows. Learning models that are constructed under various learning frameworks are relatively distinct from one another. The ability of DL to change features manually in an effective manner via the use of unsupervised or semi-supervised feature learning as well as hierarchical feature extraction is one of its benefits.

OBJECTIVES OF THE STUDY

1. to study of similarities in ml and dl .
2. to study of network security data set.

THE DIFFERENCES BETWEEN ML AND DL INCLUDE THE FOLLOWING

- Data dependencies. Deep learning stands out from more conventional forms of machine learning mostly due to the improved performance it offers with increasing amounts of data. Because deep learning algorithms need a huge quantity of data to grasp the data completely, their performance is not as good when the data volumes are little. This is because deep learning algorithms demand a significant number of data. On the other hand, the performance will be superior in this circumstance when the conventional machine-learning algorithm makes use of the previously defined guidelines.

- Hardware dependencies. The DL algorithm calls for a significant number of matrix operations. The graphics processing unit (GPU) is utilised extensively in several cases to improve and speed up matrix computations. Therefore, the graphics processing unit (GPU) is the piece of hardware required for the DL to function correctly. Traditional machine-learning methods are more dependent on low-performance computers without GPUs than they are on high-performance machines with GPUs.

- The processing of features. The practise of incorporating domain expertise into a feature extractor in order to simplify the analysis of the data and develop patterns that improve the performance of learning algorithms is referred to as feature processing. The processing of features takes a significant amount of time and needs specialised skills. When working with ML, the majority of an application's attributes need to be specified by an expert, who will then encrypt them as a data type. Pixel values, forms, textures, locations, and orientations are all examples of features that may be found in an image. The precision of the characteristics that are extracted is critical to the success of the majority of machine learning algorithms. The primary distinction that can be made between DL and more conventional machine learning methods is that DL makes an effort to derive high-level characteristics directly from the input. Consequently, DL reduces the amount of work required to create a feature extractor for every challenge.

- Problem-solving approach. When trying to solve a problem using classic machine-learning algorithms, the issue is traditionally segmented into a number of smaller problems, each of which

is then solved individually before the standard machine learning algorithm attempts to solve the larger problem again. Deep learning, on the other hand, encourages direct issue solution from beginning to finish.

- Execution time. Because there are numerous parameters in a DL algorithm, the training process often takes a longer amount of time than other types of algorithms. As a result, training a DL algorithm typically takes a considerable amount of time. While it takes precisely two weeks to finish a training session using the most powerful DL algorithm, such as ResNet, training with ML may be completed in a very short amount of time—only seconds to hours. On the other hand, the time allotted for the exam is completely different. During the testing phase, running deep learning algorithms takes a relatively short amount of time. When compared to other machine learning techniques, the testing time grows proportionally with the quantity of data. However, due to the fact that certain machine learning algorithms have very short test durations, this particular statement does not apply to all ML methods.

- Interpretability. When comparing ML with DL, interpretability is a crucially significant feature to take into consideration. DL's performance in the identification of handwritten numbers can get close to matching that of humans, which is an astounding achievement. On the other hand, a DL algorithm will not explain the reasoning behind why it produces this outcome. Naturally, from a mathematical point of view, the activation of a node inside a deep neural network is something that should be expected. But how exactly should neurons be portrayed, and how exactly should the different layers of neurons cooperate with one another? As a consequence of this, it is challenging to describe how the product was produced. On the other hand, the algorithm for machine learning offers clear criteria for why the algorithm decides thus; as a result, it is simple to explain the thinking that behind the choice.

An ML method primarily includes the following four steps :

1. Functional Engineering of Features Optional input as a foundation for forecasting (attributes, features).

2. Decide which algorithm for machine learning will best suit your needs. (Such as a classification method or a regression procedure, either with a high degree of complexity or quickly)

3. Train and assess model performance. (For each of the available algorithms, determine which model performs the best, and choose it.)

4. Apply the learned model to categorise or make predictions about the data that is not known.

The phases of a DL strategy are comparable to those of an ML approach; however, as was indicated earlier, in contrast to the techniques of machine learning, the feature extraction in DL approaches is performed automatically rather than manually. Model selection is an ongoing process of trial and error that necessitates the use of an appropriate ML / DL algorithm for each of the several sorts of missions. Approaches to Machine Learning and Deep Learning may be classified as either supervised, unsupervised, or semi-supervised. Each instance in supervised learning consists of an input sample as well as a label for that sample. The training data are analysed by the supervised learning algorithm, and the outcomes of the

analysis are used for mapping fresh instances. Unsupervised learning is a task in machine learning that derives the description of hidden structures from unlabeled data. This task requires the data to be completely unlabeled. Due to the fact that the sample has not been labelled, it is not possible to assess the precision of the algorithm's output; instead, one can only summarise and explain the most important aspects of the data. Learning that is semi-supervised is a method that combines learning that is supervised with learning that is not supervised. When learning to recognise patterns from labelled data, semi-supervised learning makes extensive use of unlabeled data in addition to the labelled data. It is possible to minimise the amount of labelling work while still reaching a high level of accuracy by using semi-supervised learning.

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)

The following are the categories that may be used to classify the outcomes of a binary classification, as indicated in Table 1:

- True Positive (TP): Positive samples that have been correctly classified by the model;
- False Negative (FN): A positive sample that has been incorrectly classified by the model;
- False Positive (FP): A negative sample that has been incorrectly classified by the model;
- True Negative (TN): Negative samples that have been correctly classified by the model; and
- False Positive (FP): A positive sample that has been incorrectly classified as a negative.

In addition, the confusion matrix may be used to produce the following metrics:

- Accuracy is calculated as follows: (TP + TN)/(TP + TN + FP + FN). A particular test data set will have a ratio that represents the proportion of samples that were properly categorised in comparison to the total number of samples. If the classes are not balanced, then this metric is not particularly informative; however, when they are, it is an excellent measure.
- Accuracy is denoted by the formula TP divided by (TP plus FP). It determines the proportion of "properly identified objects" to "actually detected things" by performing the calculation.
- The sensitivity, recall, or true positive rate (TPR) is calculated as follows: TP / (TP + FN). It does this by calculating the ratio of all "things that should be detected" to all "items that have been successfully detected."

- False Negative Rate (FNR): TP minus FN divided by FN. The proportion of incorrectly identified positive samples to total positive samples in relation to the total number of positive samples.

- The False Positive Rate, often known as the FPR, is calculated by dividing the number of true positives by the total number of true positives and false positives. The percentage of all negative samples that were incorrectly labelled as positive relative to the total number of negative samples.

- True Negative Rate, often known as TNR, is TN divided by (TN + FN). The proportion of negative samples that were accurately identified as being negative, relative to the total number of negative samples.

- The F1 score is determined by dividing the total points scored by two by the sum of the total points scored plus the first and last place finishes. It performs a calculation to get the harmonic mean of the accuracy and recall values.

- ROC: In the ROC space, the FPR is used as the abscissa for each point, while the TPR is used as the ordinate. This reflects the trade-off that the classifier makes between TP and FP. The ROC curve is a curve generated in ROC space that is used as the primary analytical tool for ROC.

- AUC: The magnitude of the area under the ROC curve is equal to the value of the AUC. The area under the curve (AUC) may have values anywhere from 0.5 to 1.0, with bigger AUCs indicating greater performance.

Precision, recall, and F1-score are three metrics that are often used in the process of developing assessment models in the field of cybersecurity. The accuracy and recall of model testing should be as high and as good as possible for the best results; yet, in practise, these two metrics may often be at odds with one another and must be carefully balanced in accordance with the requirements of the work at hand. When taking into account the outcomes of both accuracy and recall, the F1-score is the harmonic average of the two. In general, the model will perform better, and the F1-score will reflect this, the higher it is.

NETWORK SECURITY DATA SET

Data are the foundation upon which research on the protection of computer networks is built. The appropriate selection of data and the intelligent use of that data are the criteria for carrying out relevant research on security. The training effects of ML and DL models are also impacted by the quantity of the dataset that is being used. Data on the safety of computer networks may often be gathered in one of two ways: 1) directly, or 2) by using an already existing public dataset. Direct access refers to the use of a variety of methods for the direct gathering of the necessary digital data, such as the utilisation of the software tools Win Dump or Wireshark to capture network packets. This strategy is highly focused and suited for collecting short-term or small quantities of data; but, for collecting data over a longer period of time or for collecting big amounts of data, collection time and storage costs will increase. The utilisation of preexisting network security datasets may reduce the amount of time spent collecting data, which in turn can boost the effectiveness of research by allowing researchers to more rapidly get the different types of data needed for their investigations. This part will allow section IV of the study findings based on a more thorough understanding based on the introduction of some of the Security datasets that are available on the Internet..

CONCLUSION

The purpose of this research is to conduct a literature review on ML and DL techniques for the purpose of network security. This article presents the most recent uses of machine learning and deep learning in the area of intrusion detection. The primary emphasis of the article is on the most recent three years. Regrettably, there has not yet been research done to determine which form of intrusion detection is the most efficient. The examination of comparisons between the different approaches reveals that each technique of putting in place an intrusion detection system has its own set of benefits and drawbacks; this is something that is obvious from the nature of the comparisons. Because of this, it is difficult to decide which approach to the implementation of an intrusion detection system is superior to the others. When it comes to training and testing systems, having datasets for network intrusion detection is quite critical. Both the ML and DL algorithms are useless in the absence of representative data, and it is not only challenging but also time-consuming to acquire such a dataset. On the other hand, there are a lot of issues with the current population. Since network information is updated extremely frequently, deep learning and machine learning models are challenging to train and employ. Because of this, models need to be retrained quickly and over the long term. The study of this subject will thus centre its attention in the future on learning on a more gradual scale and learning that continues throughout one's life.

REFERENCES

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

[2]. 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]. 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]. 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]. 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]. 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.

[11]. R. G. Smith and J. Eckroth, "Building AI Applications: Yesterday, Today, and Tomorrow," *Ai Mag.*, vol. 38, no. 1, pp. 6–22, 2017.

[12]. P. Louridas and C. Ebert, "Machine Learning," *IEEE Softw.*, vol. 33, no. 5, pp. 110–115, 2016.