Asian Journal of Applied Science and Technology (AJAST)

(Peer Reviewed Quarterly International Journal) Volume 3, Issue 1, Pages 316-325, Jan-March 2019

A Survey of Hybrid Methods for Cyber Analytics to Assist the Intrusion Detection

Shajila V

M.Sc. Computer Science, The Center for Computer and Information Technology, University of Calicut, Thenjippalam, Malappuram, Kerala-673635

Article Received: 30 August 2018 Article Accepted: 29 December 2018 Article Published: 31 March 2019

ABSTRACT

Due to the fast growth and tradition of the internet over the last decades, the network security problems are increasing vigorously. Humans cannot handle the speed of processes and the huge amount of data required to handle network anomalies. Therefore, it needs substantial automation in both speed and accuracy. Intrusion Detection System is one of the approaches to recognize illegal access and rare attacks to secure networks. In the present world huge amounts of data are stored and transferred from one location to another. The data when transferred or stored is primed exposed to attack. Although various techniques or applications are available to protect data, loopholes exist. Thus to analyze data and to determine various kind of attack data mining techniques have emerged to make it less vulnerable. Anomaly detection uses these data mining techniques to detect the surprising behaviour hidden within data increasing the chances of being intruded or attacked. A Network Intrusion Detection System (NIDS) helps system administrators to detect network security breaches in their organizations. However, many challenges arise while developing a flexible and efficient NIDS for unforeseen and unpredictable attacks. Various hybrid approaches have also been made in order to detect known and unknown attacks more accurately. This survey work describes a focused literature survey of machine learning (ML) and deep learning (DL) methods for cyber analytics in support of intrusion detection.

Keywords: Intrusion Detection System, Machine Learning, Deep Learning, Security.

1. INTRODUCTION

Today, internet plays a significant role in our professional and personal life. It is considered as the best foundation of information about the world. As some devices connected to the internet are increasing day-by-day, so internet security threats are also increasing. Having network access by using the internet, attackers have some ways to steal, destroy, or gain unauthorized data. The question is how to make our networks secure from such threats. Intrusion Detection System (IDS) is designed for this purpose.

The Internet has become the most essential tool and one of the best sources of information about the current world. Internet can be considered as one of the major components of education and business purpose [6]. Therefore, the data across the Internet must be secure. Internet security is one of the major concerns now-a-days. As Internet is threatened by various attacks it is very essential to design a system to protect those data, as well as the users using those data. Intrusion detection system (IDS) is therefore an invention to fulfill that requirement. Network administrators adapt intrusion detection system in order to prevent malicious attacks. Therefore, intrusion detection system became an essential part of the security management. Intrusion detection system detects and reports any intrusion attempts or misuse on the network. IDS can detect and block malicious attacks on the network, retain the performance normal during any malicious outbreak, perform an experienced security analysis.

The basic purpose of IDS is to assist networks in coping with the network attacks so that they can identify anomalous, wrong and unsuitable actions of a computer system [3].

CITE THIS ARTICLE: Shajila V., "A Survey of Hybrid Methods for Cyber Analytics to Assist the Intrusion Detection ", *Asian Journal of Applied science and Technology,* Volume 3, Issue 1, Pages 316-325, Jan-March 2019

To protect systems from harmful attacks, network administrators use IDS. For security supervision, IDS became a crucial part. To hold operation normal throughout the harmful attack, intrusion detection system can identify and block harmful outbreaks [5]. Intrusion may be defined as an activity that essentially attempts to compromise integrity, authenticity, confidentiality, availability of system resources. For an organization to be safe, efficient and most reliable it must maintain several layers of security. These include network security and information security. This challenge is becoming much more complex currently as systems and services are becoming complex facilitating several new possibilities for attackers. The components of intrusion detection system are depicted in figure 1.

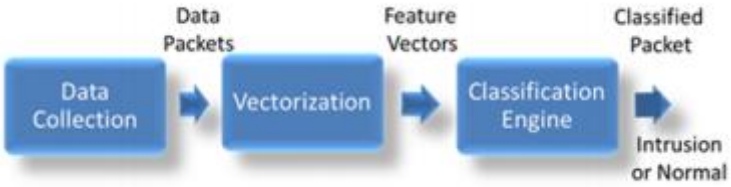

Figure No: 1 Constituents of Intrusion Detection System

• Firstly, there should be a data collection mechanism which tracing the network flows

• Then, these data need to be used to identify the features, and a feature vector needs to be created.

• Finally, with the use of this vector, a classification engine is executed, and the traced flow is identified either as normal or as an intrusion according to previous knowledge.

Intrusion detection system approaches can be classified in 2 different categories. One of them is anomaly detection and the other one is signature based detection, also known as misuse detection based detection approach. The misuse detection is used to identify attacks in a form of signature or pattern. As misuse detection uses the known pattern to detect attacks the main disadvantage is that it will fail to identify any unknown attacks to the network or system. On the other hand, anomaly detection is used to detect unknown attacks. There are different ways to find out the anomalies.

Hence different types of machine learning and deep learning techniques are introduced in order to identify the anomalies. The work provides the survey of various above described methods. The accuracy level of all the methods are compared and demonstrated.

2. RELATED WORK

Nguyen et al. [1] describe ML techniques for Internet traffic classification. The techniques described therein do not rely on well-known port numbers but on statistical traffic characteristics. Their survey only covers papers published in 2004 to 2007, where our survey includes more recent papers. Unlike Nguyen et al. [1], this paper presents methods that work on any type of cyber data, not only Internet Protocol (IP) flows.

Teodoro et al. [2] focus on anomaly-based network intrusion techniques. The authors present statistical, knowledge-based, and machine-learning approaches, but their study does not present a full set of state-of-the-art machine-learning methods. In contrast, this paper describes not only anomaly detection but also signature-based methods. Our paper also includes the methods for recognition of type of the attack (misuse) and for detection of an attack (intrusion). Lastly, our paper presents the full and latest list of ML/DM methods that are applied to cyber security.

Sperotto et al. [3] focus on Network Flow (NetFlow) data and point out that the packet processing may not be possible at the streaming speeds due to the amount of traffic. They describe a broad set of methods to detect anomalous traffic (possible attack) and misuse. However, unlike our paper, they do not include explanations of the technical details of the individual methods.

Wu et al. [4] focus on Computational Intelligence methods and their applications to intrusion detection. Methods such as Artificial Neural Networks (ANNs), Fuzzy Systems, Evolutionary Computation, Artificial Immune Systems, and Swarm Intelligence are described in great detail. Because only Computational Intelligence methods are described, major ML/DM methods such as clustering, decision trees, and rule mining (that this paper addresses) are not included.

3. MACHINE LEARNING APPROACHES

Signatures based IDSs rely on humans to create, test, and deploy the signatures. Thus, it may take hours or days to generate a new signature for an attack, which can be too long when dealing with rapid attacks. Nevertheless, in order to offer a human-independent solution to the abovementioned problem, anomaly based IDSs based on machine learning techniques provide an added advantage. Anomaly based IDSs using machine learning techniques have the ability to implement a system that can learn from data (experience) and give the decision for unseen data [7].

i) Neural Networks

Artificial neural networks consists of a collection of processing elements that are highly interconnected and transform a set of inputs to a set of desired outputs that is inspired by the way biological nervous systems, such as the brain, process information. The technique of Neural Networks NN follows the same theories of how the human brains works. The Multilayer Perceptions MLP has been widely used neural network for intrusion detection. They are capable of approximating to arbitrary accuracy, any continuous function as long as they contain enough hidden units. This means that such models can form any classification decision boundary in feature space and thus act as non-linear discriminate function. When the NN is used for packets classification, there is one input node for each element of the feature vector. There is usually one output node for each class to which a feature may be assigned (shown in Fig. 2). The hidden nodes are connected to input nodes and some initial weight assigned to these connection. These weights are adjusted during the training process. MPL NN Structure One learning algorithm used for MLP is called back-propagation rule. This is a gradient descent method and based on an error function that represents the difference between the network's calculated output and the desired output. This error function is

defined, based on the Mean Squared Error MSE. The MSE can be summed over the entire training set [8]. In order to successfully learn, the network's true output should approach the desired output by continuously reducing the value of this error. The back propagation rule calculates the error for a particular input and then back- Propagates the error from one layer to the previous one.

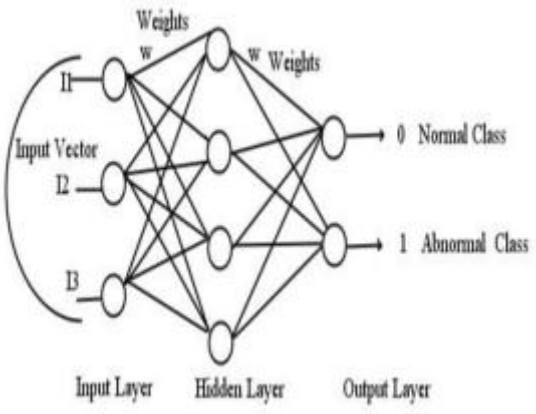

Figure No: 2 Simple Framework of ANN

The connection weights, between the nodes, are adjusted according to the back-propagated error so that the error is reduced and the network learns. The input, output, and hidden layers neurons are variable. The neural network intrusion detection system consists of three phases:

- Using automated parsers to process the raw TCP/IP dump data in to machine-readable form or use some benchmark dataset contain parsed connection records.
- Training: neural network is trained on different types of attacks and normal data.
- Testing: performed on the test dataset.

Neural network based intrusion detection system, intended to classify the normal and attack patterns and the type of the attack. It is observed that the long training time of the neural network was mostly due to the huge number of training vectors of computation facilities. However, when the neural network parameters were determined by training, classification of a single record was done in a negligible time. Therefore, the neural network based IDS can operate as an online classifier for the attack types that it has been trained for. The only factor that makes the neural network off-line is the time used for gathering information necessary to compute the features.

ii) Decision Tree

Decision Tree algorithm is normally used for classification problem. In this algorithm, the data set is learnt and modeled. Therefore, whenever a new data item is given for classification, it will be classified accordingly learned from the previous dataset. Decision Tree algorithm can also be used for Intrusion Detection. For this reason, the algorithm will also learn and models data based on the training data. As a result, the model can classify which attack types does a future data belongs to based on the model built. One of the strength of Decision Tree is it can works well with huge data sets. This is important as large amount of data flow can be found across the computer networks [8]. In real-time Intrusion Detection, it works well because Decision Tree gives the highest detection performance

and can construct and interpret model easily. Another useful property of Decision Tree in Intrusion Detection model is its generalization accuracy. This is due to the trends in the future where there will always be some new attacks, and by having generalized accuracy provided by Decision Tree, these attacks can be detected.

B. Deep Learning

Deep learning is an improved machine learning technique for feature extraction, perception and learning of machines. Deep learning algorithms perform their operations using multiple consecutive layers. The layers are interlinked and each layer receives the output of the previous layer as input. It is a great advantage to use efficient algorithms for extracting hierarchical features that best represent data rather than manual features in deep learning methods.

The term deep learning comes from the advancements of neural network. In deep learning, various methods have applied in order to overcome the limitations of the hidden layer. Basically, those methods employ consecutive hidden layers which hierarchically structured. Since a lot of methods belong to deep learning method, the classification of each deep learning method is essential [9]. The deep learning is classified into three sub-groups, generative, discriminative and hybrid. The classification is based on the intention of architectures and techniques, e.g., synthesis/generation or recognition/classification. The classification of the deep learning methods is shown in Fig. 3.

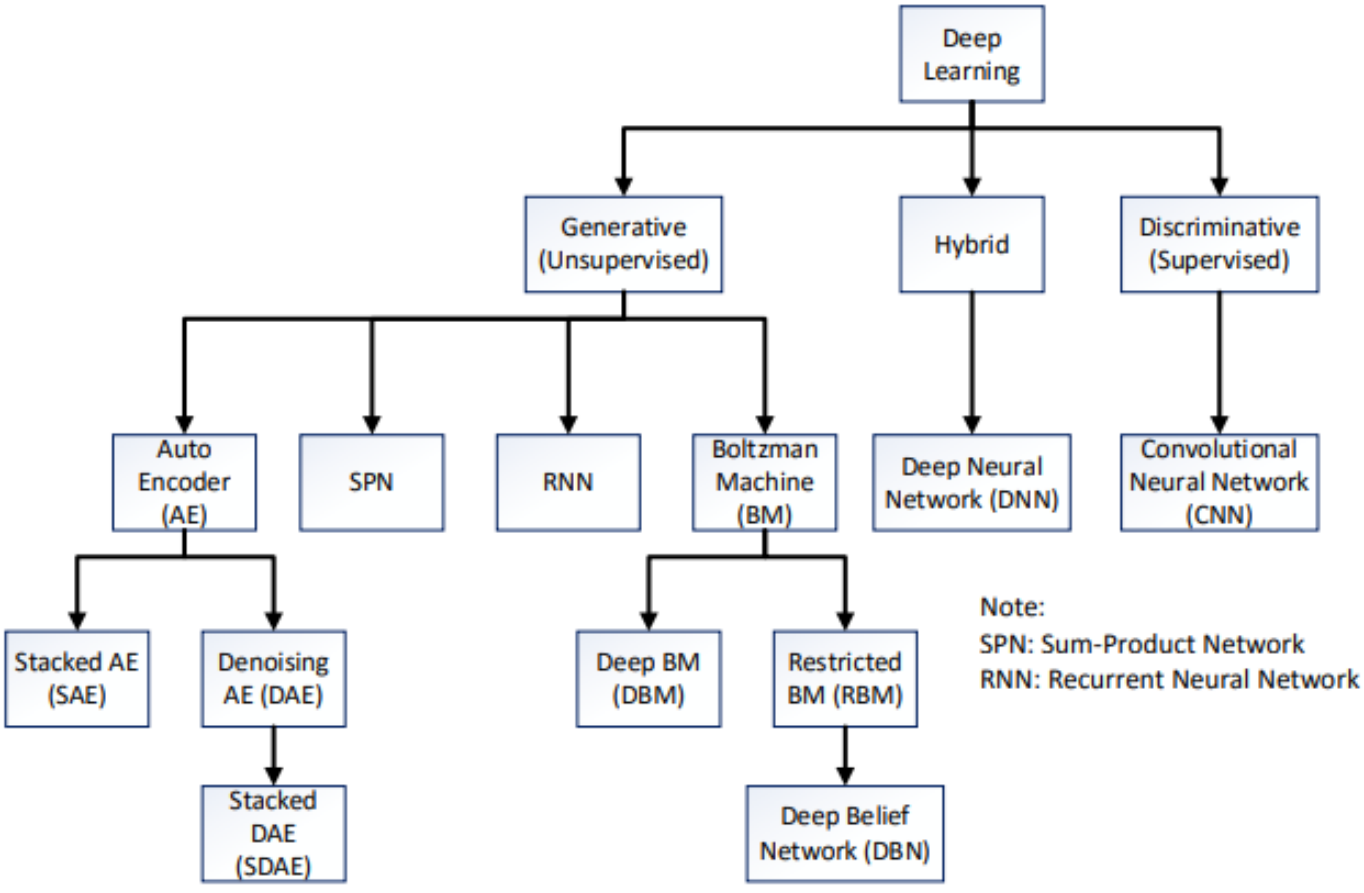

Figure No: 3 Categorization of Deep Learning Algorithms

 Online ISSN: 2456-883X

i) Unsupervised Learning

Unsupervised learning or so called generative architectures, which the data are unlabeled. The main concept of applying generative architectures to pattern recognition is unsupervised learning or pre-training. Since learning the lower levels of subsequent networks are difficult, deep generative architectures are needed. Thus, with limited training data, learning each lower layer in layer-by-layer approach without relying on all the layers above is important.

There are a number of methods that classified as unsupervised learning as follows: Basically, Auto Encoder (AE) is just similar to Artificial Neural Network (ANN) which is the number of hidden layer is three only. The main different is the nodes in the input layer are the same with the output layer. Meanwhile, the nodes in middle hidden layer are representing new features set which are low-dimensional. This architecture leads to an ability that can reconstruct the data after complicated computations. AE aims to learn a compact set of data efficiently. AE structures can be stacked to build deep networks. Each training results of the middle layer are cascaded. This structure is called Stacked Auto-Encoder (SAE) which can learn a lot of new features in different depths. In order to train more precisely, we can append an additional layer with labels once we have large amount of tagged samples. In addition, a Denoising Auto Encoder (DAE) is trained to reconstruct a clear correction input from a corrupted by noise input. The DAE may be also stacked in order to build deep networks too.

BM is a network of binary units that symmetrically paired. BM has neuron units' structure that makes stochastic decisions about whether active or not. If one BM result is cascaded into multiple BMs, it's called Deep BM (DBM). Meanwhile, Restricted Boltzmann machine (RBM) is a customized BM without connections among the hidden units. RBM consists of visible and hidden variables such that their relations can be figured out. If multiple layers are stacked, layer-by-layer scheme, called as Deep Belief Network (DBN). DBN could be used as a feature extraction method for dimensionality reduction when unlabeled dataset and back-propagation are used (which means unsupervised training). In contrast, DBN is used for classification when appropriate labeled dataset with feature vectors are used (which means supervised training).

ii) Supervised Learning

Supervised learning or discriminative deep architectures will be able to distinguish some parts of data for pattern classification. An example of discriminative architecture is CNN which employs a special architecture particularly suitable for image recognition. The advantage of CNN is fast to train because of its structure. CNN can train multilayer networks with gradient descent to learn complex, high-dimensional, nonlinear mappings from large collections of data. CNN uses three basic concepts: local receptive fields, shared weights, and pooling. One extensive research that successfully deployed using CNN is AlphaGo by Google.

iii) Hybrid Method

Hybrid deep architectures combine both generative and discriminative architectures. Basically, the hybrid architecture aims to distinguish data as well as discriminative approach. However, in the early step, it has assisted in a significant way with the generative architectures results. An example of hybrid architecture is Deep Neural

Network (DNN). However, there often confusion terms between DNN and DBN. In the open literatures, DBN also uses back propagation discriminative training as a "fine-tuning". This concept of DBN is really similar to Deep Neural Network (DNN). DNN is defined as a multilayer network with cascaded fully connected hidden layers, and is often use stacked RBM as a pre-training phase.

4. RESULTS AND DISCUSSION

The performance of the machine learning and deep learning algorithms are tested based on the following metrics:

• **Accuracy:** Defined as the percentage of correctly classified records over the total number of records.

• **Precision (P):** Defined as the % ratio of the number of true positives (TP) records divided by the sum of true positives (TP) and false positives (FP) classified records.

$$P = \frac{TP}{(TP + FP)} \times 100\%$$

• **Recall (R):** Defined as the % ratio of number of true positives records divided by the sum of true positives and false negatives (FN) classified records.

$$R = \frac{TP}{(TP + FN)} \times 100\%$$

• **F-Measure (F):** Defined as the harmonic mean of precision and recall and represents a balance between them.

$$F = \frac{2.P.R}{(P + R)}$$

In this work, the performance of machine learning and deep learning are compared. Table 1 and figure 4 denote the correctly classified and incorrectly classified instances for both the algorithms.

S.No	Algorithms	Correctly Classified Instances	Incorrectly Classified Instances
1	Machine Learning	94.07%	5.83%
2	Deep Learning	95.17%	4.83%

Table No: 1 Comparison Based on Classified Instances

Machine learning techniques have ability to implement a system that can learn from data. For example, a machine learning system could be trained on incoming packets to learn to distinguish between intrusive and normal packet. After learning, it can then be used to classify new incoming packets into intrusive and normal packets.

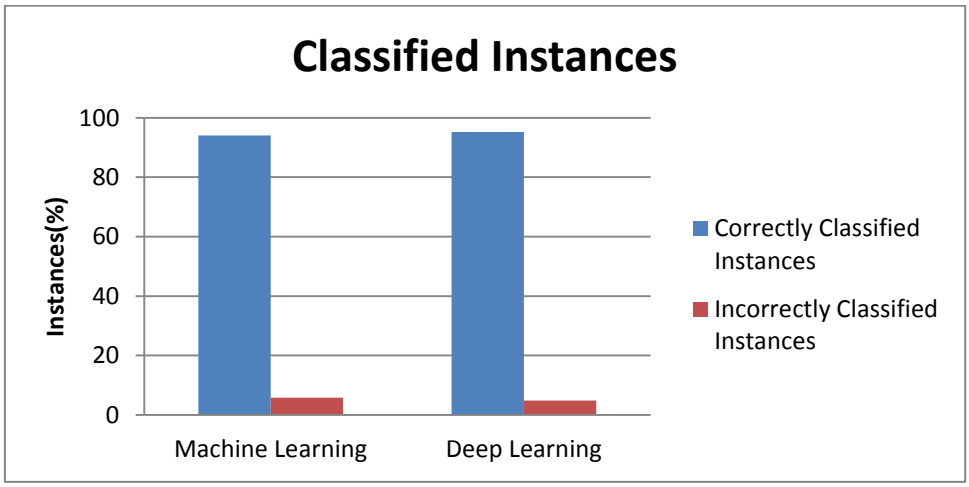

Figure No: 4 Comparison Based on Classified Instances

In the same way, the accuracy, precision, detection rate and F-measure are analyzed and compared in table 2 and figure 5.

S.No	Algorithms	Precision	Recall	F-Measure
1	Machine Learning	0.953	0.952	0.952
2	Deep Learning	0.996	0.996	0.995

Table No: 2 Comparison Based on Accuracy, Precision and F-Measure

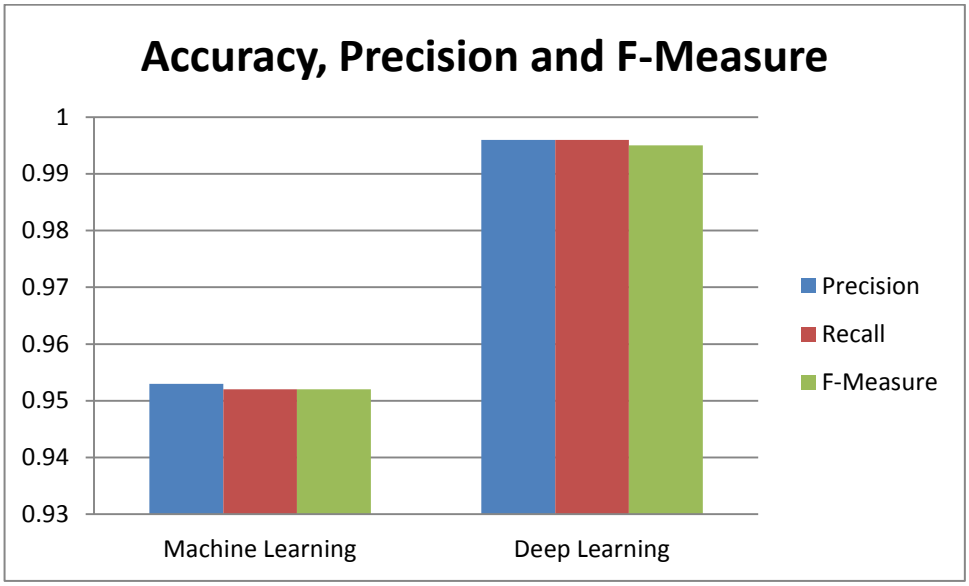

Figure No: 5 Comparison Based on Accuracy, Precision and F-Measure

The goal of deep learning methods is to learn feature hierarchies with features composed by lower level into higher level features. The methods can learn features independently at multiple levels of abstraction, and thus discover complicated functions mapping between the inputs to the output directly from raw data without depending on customized features by experts. In higher-level abstractions, humans often have no idea to see the relation and

connection from raw sensory input. Therefore, the ability to learn complex features will become necessarily needed as the amount of data increased sharply. In some literatures, this capability is often called as feature extraction or feature engineering. Feature engineering is the transforming process of raw data input into features which represents the problem properly, and thus improves model accuracy on uncovered data.

5. CONCLUSION

Machine learning for intrusion detection has received much attention in the computational intelligence community. In intrusion detection algorithm, huge volumes of audit data must be analyzed in order to construct new detection rules for increasing number of novel attacks in high speed network. Intrusion detection algorithm should consider the complex properties of attack behaviors to improve the detection speed and detection accuracy. Analyze the large volume of network dataset and improve the performance of detection accuracy, intrusion detection become an important research field for machine learning. Advantage of machine learning-based detection systems detect or categorize persistent features without any feedback from the environment. Disadvantages of learning-based detection systems are if a credible amount of normal traffic data is not available, the training of the techniques becomes very difficult. Deep learning is a class of machine learning methods, where exploits the cascaded layers of data processing stages in hierarchical structure for unsupervised feature learning and for pattern classification. The principle of deep learning is to process hierarchical features of the provided input data, where the higher-level features are composed by lower-level features. Furthermore, the deep learning method can integrate a feature extractor and classifier into one framework which learns feature representations from unlabeled data autonomously, and thus the security experts doesn't need to craft the desired features manually. Essentially, deep learning methods can discover sophisticated underlying structure/feature from abstract aspects. This abstraction ability of deep learning makes it feasible to abstract benign or malicious features among the provided data.

6. REFERENCES

[1] T. T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," IEEE Commun. Surv. Tuts., vol. 10, no. 4, pp. 56–76, Fourth Quart. 2008.

[2] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," Comput. Secur., vol. 28, no. 1, pp. 18–28, 2009.

[3] A. Sperotto, G. Schaffrath, R. Sadre, C. Morariu, A. Pras, and B. Stiller, "An overview of IP flow-based intrusion detection," IEEE Commun. Surv. Tuts., vol. 12, no. 3, pp. 343–356, Third Quart. 2010.

[4] S. X. Wu and W. Banzhaf, "The use of computational intelligence in intrusion detection systems: A review," Appl. Soft Comput., vol. 10, no. 1, pp. 1–35, 2010.

[5] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization." in ICISSP, 2018, pp. 108–116.

[6] Salama, Mostafa A, EID, Heba F, Ramadan, Rabie A, Darwish, Ashraf, "Hybrid Intelligent Intrusion Detection Scheme", In: Springer Berlin Heidelberg (2011), S. 293–303.

[7] N. F. Haq, A. R. Onik, A. K. Hridoy, M. Rafni, F. M. Shah, and D. Farid, "Application of Machine Learning Approaches in Intrusion Detection System: A Survey," Int. J. Adv. Res. Artif. Intell., vol. 4, no. 3, pp. 9–18, 2015.

[8] A. Osareh and B. Shadgar, "Intrusion Detection in Computer Networks based on Machine Learning Algorithms," Ijcsns, vol. 8, no. 11, p. 15, 2008.

[9] U. Cekmez, Z. Erdem, A. G. Yavuz, O. K. Sahingoz, and A. Buldu, "Network anomaly detection with deep learning," in 2018 26th Signal Processing and Communications Applications Conference (SIU), May 2018, pp. 1–4.