International Journal of Computer Engineering in Research Trends
Multidisciplinary, Open Access, Peer-Reviewed and fully refereed
Survey Paper Volume-7, Issue-12, December 2020 Regular Edition E-ISSN: 2349-7084

Cloud Security Solutions through Machine Learning- Approaches: A Survey

[1]K. Samunnisa, [2]G. Sunil Vijaya Kumar, [3]K. Madhavi
[1]. *Research Scholar, Department of Computer Science and Engineering, JNTUA College of Engineering, Ananthapur, A.P .*
[2] *Professor, Department of Computer Science and Engineering, Global college of engineering and technology, Kadapa.*
[3]*Professor, Department of Computer Science and Engineering, JNTUA College of Engineering, Ananthapur, A.P*

Available online at: http://www.ijcert.org

Received: 03/12/2020 *Revised: 17/12/2020* *Accepted: 18/12/2020* *Published: 02/01/2021*

Abstract: Cloud computing provides services to the consumer such as large-scale computation, data storage, virtualization, high expansibility, high reliability and low price service. The storage of data in cloud is more sensitive to users as it is stored in the third party storage and it is one of the major problems identified. Data protection consists of a variety of regulations, protocols, processes and techniques, which operate together to secure cloud-based services, networks and data. All these protection mechanisms are intended to secure data, secure customer privacy, and promote compliance with regulations and set system and user authentication guidelines. Cloud security issues are DDoS threats, violations of privacy, lack of privacy and insecure access points. Researchers have performed an investigation into many intrusion prevention methods for cloud infrastructure detection of intrusion. Most of them address conventional intrusion and anomaly detection strategies and concentrate on best practice in cloud protection, such as server & virtualization security, host & middleware security and device and data security. This paper focuses on machine learning-based cloud computing solutions, and the current study is focused on technology issues. Related open problems are defined as being of potential significance.

Keywords: Anomaly detection, Cloud Storage, cloud computing, machine learning techniques, Supervised- and Unsupervised-learning.

1. Introduction

Cloud-based data storage capabilities have grown in recent years. Moreover, due to reliability and cost savings, businesses are transferring their data to these servers. It's also used by large computing capacity in non-traditional industries such as online gaming and social media. By 2024, the worldwide demand for cloud protection will hit USD 12.64 billion, powered by the growing usage of cloud computing for data storage and cyber threat advances [1]. It may be a device or service that focuses on regulatory regulation, governance and data protection. Gartner expects that 95% of cloud failures would be the responsibility of the company. With approximately 70% of all cloud-based businesses, cloud protection risks should affect any enterprise. Like many other technical technologies, cloud computing has offered many opportunities. For example, it allowed a vast volume of data and different resources to be stored. This initiative has addressed the issue of resource limitations and lowered transaction costs by exchanging scarce services with different users. The cloud offers a modular platform for seamless Information control, online access, accessibility and cost-effectiveness. Reliability and efficiency of services require stable mechanisms against security threats [2]. However, as more mission-critical technologies move into the cloud, there are growing questions regarding data protection and cyber protection.

Cloud computing has been one of the essential subjects of technical analysis of recent years. These studies cover the protection of data storage, network security, and device security. The NIST describes cloud computing as [3] a "platform for flexible resource pooling, universal, on-demand access that can be

© 2020, IJCERT All Rights Reserved

conveniently provided through various forms of service provider interaction."

The cloud computing protocol incorporates Pay as You Go (PAYG), where allows users paid for the software they are using. The PAYG model gives consumers the option to tailor client and end-user applications, data, production tools and machine resources. These advantages are the reason why the science community has invested a lot of work into the cutting-edge concept [4].

There are many facets of cloud computing, including the paradigm for cloud architecture that offers a specific form of cloud infrastructure that is defined by its complexity, ownership and usability. Cloud storage is essentially allowed by exchanging services between individual computers or local servers. The purpose and essence of the cloud are related to the architecture model. The implementation model consists of three types: public cloud, private cloud and hybrid [5].

Networks have a growing influence on everyday lives with the reasons mentioned above, making data security a significant area of study. In specific, cyber-security strategies include anti-virus applications, firewalls and IDSs. These strategies avoid internal and external threats by networks. An IDS is a type of identification system that plays a crucial role in protecting cybersecurity by tracking the states of a network's software and hardware.

There have been several mature IDS products. Many ISDs, however, also suffer from high false alarm rates, which produce several alarms for low non-dangerous conditions that increase the safety analysts' workload and can cause severe hazardous attacks to be missed. Many researchers have, therefore based their efforts on improving IDSs with better detection rates and lower fake alarm rates. Another problem with current IDSs is that unexplained threats cannot be identified. As network dynamics are evolving rapidly, attack types and new threats are continuously emerging. Therefore, IDSs that can detect unknown attacks must be created.

Researchers have started to concentrate on developing IDSs using machine learning approaches to resolve the above problems. ML is a kind of artificial intelligence technology that can find useful information automatically in large datasets [6]. Machine-based IDS can reach sufficient detection levels if appropriate training data are usable, and machine-based learning models are versatile sufficiently to identify attack variants and new attacks. Moreover, machine-to-learn IDSs are not highly backed by domain information, so they are easy to develop and create, and ML techniques can achieve excellent performance.

The goal of this survey is to identify and sum up the IDSs presented to date in the field of machine learning, to abstract the principal ideas of applying machine learning to security issues and to examine existing problems and possible innovations. We have chosen representative papers for this study, published between 2015 and 2019, which demonstrate current development. Several previous surveys[7–9] listed research activities with their algorithms for advanced machine learning. These surveys are explicitly designed to implement new machine learning algorithms for IDSs, which can allow computer researchers to understand. This type of taxonomic scheme, however, stresses real technical application rather than computer security concerns. Thus, these experiments do not deal explicitly with how IDS domain issues can be addressed using machine learning. We suggest a new data-centred IDS taxonomy in this study to resolve this issue and present the relevant studies after this taxonomy.

This paper discusses the shortcomings of previous surveys and outlines the IDS identification process and presents a detailed overview of threat models, threats, and IDS approaches in a cloud environment. Our main paper contributions can be summarized as follows.

- Discuss definitions and problems in cloud defense
- Clarify Cloud IDS Classification System Taxonomy
- Comprehensive IDS Basic Machine Learning Algorithms Investigation
- Machine Learning-based study on IDSs and their problems.

The rest of the paper is organized as follows Section 2 gives Cloud Security Categories and Issues, Section 3 Evaluation of Cloud-Based IDS, Section 4 describes Common Machine Learning Algorithms in IDS, and Section 5 explores the Research on Machine Learning-Based IDS, Section 6 presents Cloud Security challenges and Section 7 concludes the survey paper.

2. Cloud Security Categories and Issues

The stability of all data layers in public and private databases is protected by cloud encryption. Except for the application layer, protection for cloud applications includes Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) and Cloud device security and makes sure the application layer as seen in Figure 1.

© 2020, IJCERT All Rights Reserved

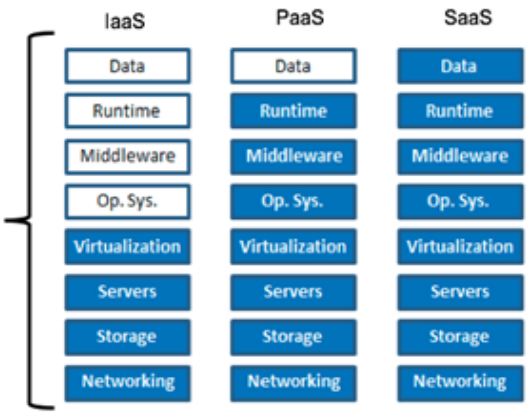

Figure 1. Cloud Security

This paper offers a crucial part of this analysis by analyzing emerging cloud protection problems and cutting-edge technology strategies. The paper defines in (Table 2) 28 security problems and categorizes into five sections (Table 1). The latest security strategies and state-of-the-art countermeasures are often related.

2.1 Categories and Issues

We classify cloud computing security related issues into the following five categories, which are also summarized in Table 1. A similar approach to classify the issues is found in [10], but it is limited to a small set of cloud security concerns and only partially covers four categories.

Category 1: Class Security Requirements deals with regulatory and regulating bodies that specify cloud protection policies to protect a cloud operating environment. It contains service level agreements, evaluations and other customers, service suppliers and other stakeholder's agreements.

Category 2: The Network category refers to the mechanism used for users to connect to the cloud networks to execute the necessary calculations. It involves plugins, network links and login knowledge sharing.

Category 3: The Access Control category is a user-oriented area that involves problems of recognition, authorization and authorization.

Category 4: The term Cloud Computing covers security problems in SaaS, PaaS and IaaS and is relevant in particular to the virtualization environment.

Category 5: The privacy division includes questions of data security and secrecy.

Table 1 Cloud Security Categories.

Category No	Category	Description
Category 1	Security Standards	Describes the standards for taking precautionary steps to deter attacks in cloud computing. It controls cloud storage policies for protection without impacting security and performance.
Category 2	Network	This includes network attacks, such as Link Failure, Service Denial (DoS), DDoS, flood attack, web protocol vulnerabilities, etc.
Category 3	Access Control	Covers security and monitoring of entry. It defines problems that concern user identity and data storage protection.
Category 4	Cloud Infrastructure	Covers cloud infrastructure-specific threats Such as tampered binaries and privileged insiders (IaaS, PaaS, and SaaS).
Category 5	Data	Covers security concerns related to data including data transfer, completeness, encryption and data protection.

© 2020, IJCERT All Rights Reserved https://doi.org/10.22362/ijcert/2020/v7/i12/v7i1203

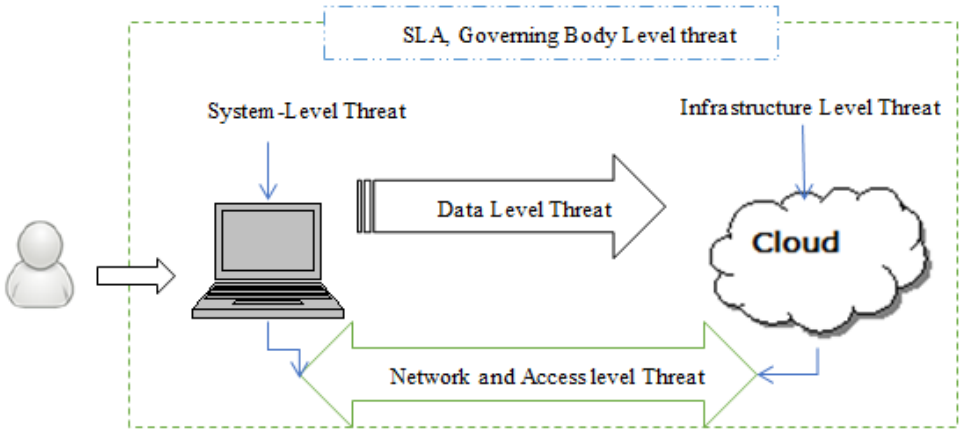

Figure 2. Cloud components that are inclined to security threats

Figure 2. Shows the elements of the cloud that can pose security concerns. -- aspect is vulnerable to security attacks such as protocols, customers, cloud infrastructure and network, which needs avoidance, identification, or response techniques for attacks. Table 2 maps the reported cloud protection problems into the relevant previously specified categories (Table 1). There is a need to pay particular attention to standard safety requirements, including Protected Sockets Layer (SSL)/Transport Layer Protection (TLS), XML Signature, XML Encryption Syntax and Encoding, and Key Management Interoperability Protocols. There are no acceptable security requirements in cloud computing [11]. While safety requirements are appropriately described, many safety concerns remain to be related to enforcement risks due to the lack of audit governance and appraisal of business quality [11]. Cloud customers lack sufficient knowledge of their supplier's protocols, processes and activities, especially in the areas of identity management and role segregation.

Table 2 Cloud Security Categories with Issues.

Category	Issues
Security Standards (C1)	• Lack of security standards • Compliance risks • Lack of auditing • Lack of legal aspects (Service level agreement) • Trust
Network (C2)	• Proper installation of network firewalls • Network security configurations • Internet protocol vulnerabilities • Internet Dependence
Access (C3)	• Account and service hijacking • Malicious insiders • Authentication mechanism • Privileged user access • Browser Security

© 2020, IJCERT All Rights Reserved https://doi.org/10.22362/ijcert/2020/v7/i12/v7i1203

Cloud Infrastructure (C4)	• Insecure interface of API • Quality of service • Sharing technical flaws • Reliability of Suppliers • Security Misconfiguration • Multi-tenancy • Server Location and Backup
Data (C5)	• Data redundancy • Data loss and leakage • Data location • Data recovery • Data privacy • Data protection • Data availability

Auditability is one of the most critical aspects of cloud computing reliability, but we do not have an audit network for cloud providers [13, 14]. When the vendor delivers a service to a third party that does not provide apparent features, consumers must be able to control the entire operation. Security standards (C1) and regulatory bodies are part of Service Level Agreements (SLA) and legislative frameworks not integrated into cloud computing activities [14, 15] respectively.

Network category (C2) is perceived to be the most significant safety threats in clouds since cloud computing is more likely than conventional computing paradigms to target networks [16]. Furthermore, cloud operations are closely related and rely heavily on networking. Cloud network security concerns are also given more significant focus in this review than other security groups.

Quality of Service (QoS) is an unattended problem [15] because many cloud service providers rely only on upon fast and low-cost performance [14]. In this job, we consider QoS in the domain of any feature or operation which affects protection directly or indirectly. A small failure in the configuration of one or more of the cloud components can have profound implications so that multiple providers can exchange system configurations [18].

There have been significant and essential problems in numerous case studies that require data to be appropriately stored, shared, secured, managed and available in time of need for data redundancy[19], data retention and leakage[20], data location[21], data replication [18], privacy [21], data protection[22], and data availability[22].

3. Evaluation of Cloud-Based Ids

Several scholars have adapted the standard IDS Method to the cloud world The European Union Agency for Network and Information Security (ENISA) has been working hard to overcome many cloud-related security issues. It gives customers knowledge that lets them identify, analyze and handle risks as they transition to the clouds. It also offers consultancy services on SLAs to maximize safety benefits. ENISA also conducts joint projects with different partners to identify core cloud services and evaluate in those situations the consequences of the cloud service failure.

3.1 Intrusion Detection Systems (IDS)

For IDS, an intruder involves an attempt to access information about computer networks or to unlawfully or unlawful harm to system operations.

© 2020, IJCERT All Rights Reserved https://doi.org/10.22362/ijcert/2020/v7/i12/v7i1203

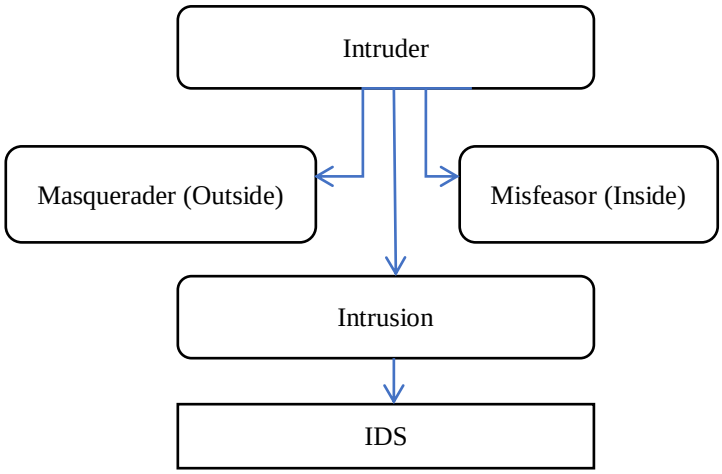

Figure 3. IDS System Structure

An IDS [23] is a data safety programme intended to identify a wide variety of flaws, from alleged compromises by external parties to device exploitation and harassment by insiders, as shown in Figure 3. IDS' essential functions are to track hosts and networks, evaluate computer system behaviours, produce warnings and respond to unusual behaviours. As IDSs are typically used to close secure network nodes (e.g. switches in major network areas), because of their control of relevant hosts and networks.

Two types of IDS classification methods exist an approach based on identification and methods based on data source. IDSs may be broken into signature-based IDS or misuse-based IDS and anomaly-based IDS identification among the identification based approaches. IDSs can be divided into host-based and network-based methods by data-source based methods [24]. This study incorporates these two types of IDS classification methods by considering the data source as the primary concern of classification and the identification system as a secondary feature. Figure 4 demonstrates the current IDS grouping. The study focuses on machine learning approaches concerning identification techniques. Besides, in Section 4, introduce how to apply machine learning in-depth to IDS using different types of data.

© 2020, IJCERT All Rights Reserved https://doi.org/10.22362/ijcert/2020/v7/i12/v7i1203

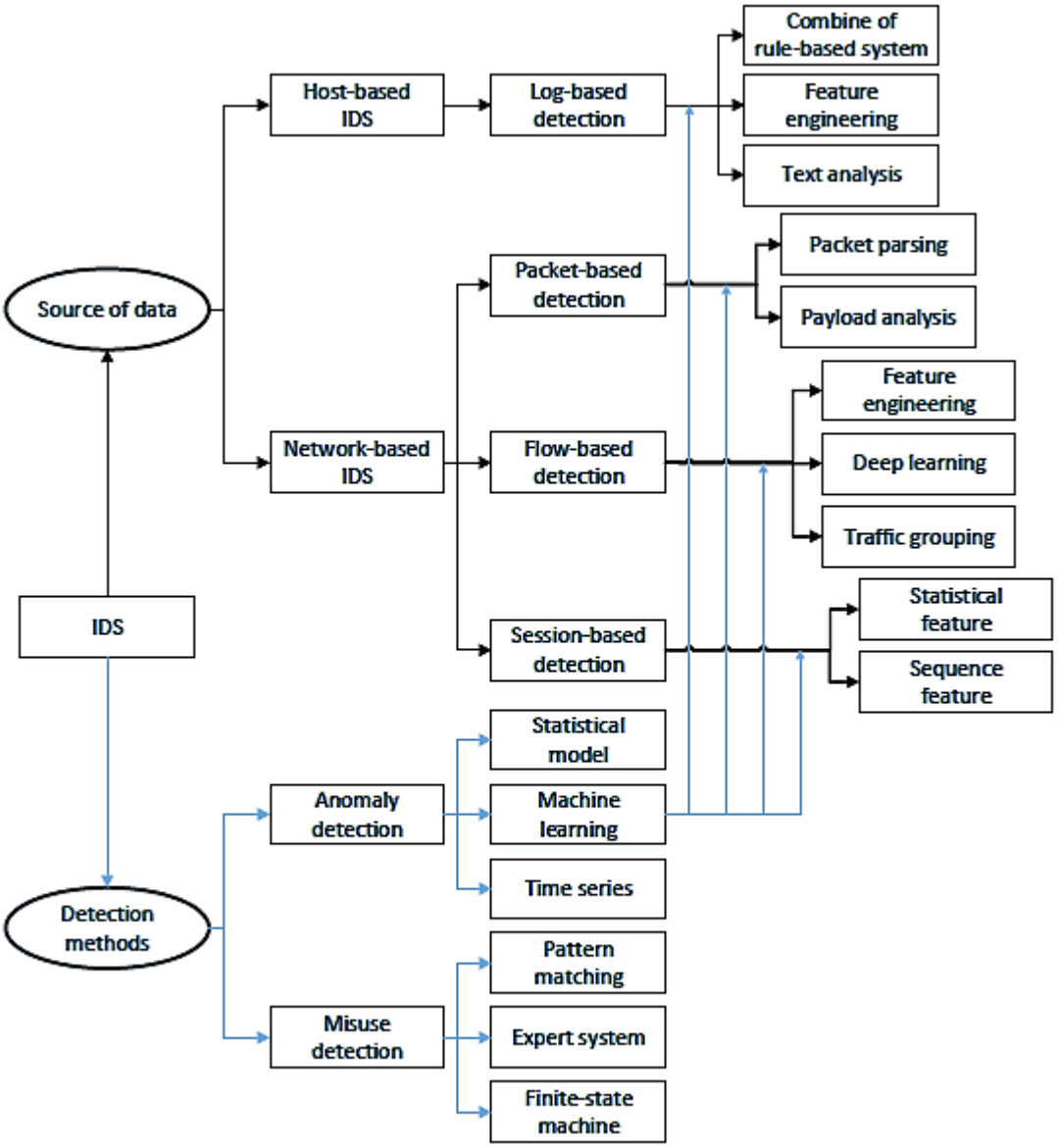

Figure 4. IDS Classification System

3.2 Classification by Detection Methods

Misuse detection is sometimes considered a signature-based detection—the underlying principle of describing attacks as signatures. The method of identification correlates with sample signatures using a signature database. The key challenge in developing frameworks for misuse identification is creating successful signatures. The benefit of abuse detection is that it has a low false alarm rate and discusses in-depth threat forms and potential reasons; the drawbacks are the high missed alert rate, the ability to track unexpected threats and the need to hold an extensive signature database.

The advantages of anomaly detection are high generalization and the ability to spot unexpected threats. However, their limitations are high false alarm rates and unable to provide any explanation for an abnormality. Table 3 lists the key distinctions between abuse detection and anomaly detection.

© 2020, IJCERT All Rights Reserved https://doi.org/10.22362/ijcert/2020/v7/i12/v7i1203

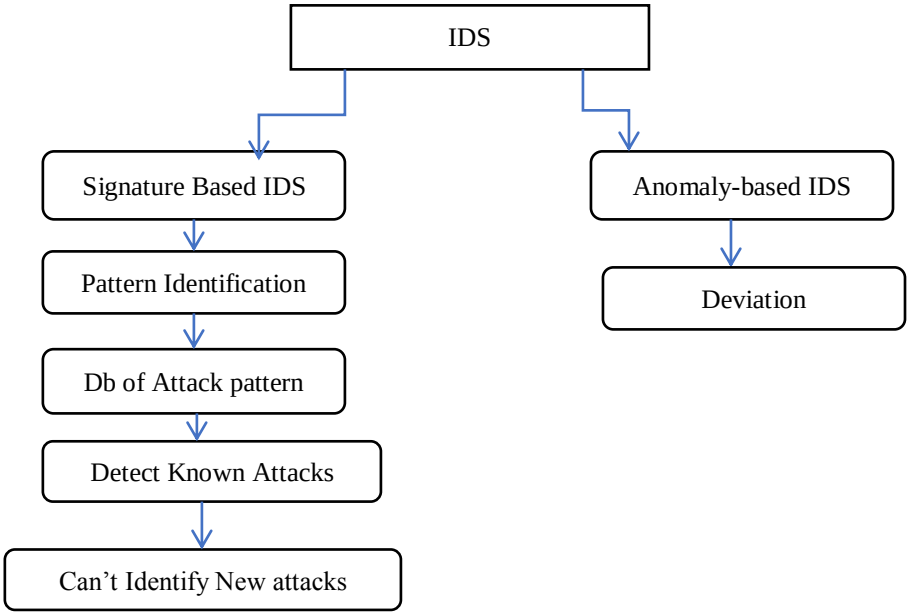

Figure 5. IDS Detection Classifications

Table 3 Difference between misuse detection and anomaly detection

	Misuse Detection	**Anomaly Detection**
Detection performance	Low false alarm rate; High missed alarm rate	Low missed alarm rate; High false alarm rate
Detection efficiency	High, decrease with scale of signature database	Dependent on model complexity
Dependence on domain knowledge	Almost all detections depend on domain knowledge	Low, only the feature design depends on domain knowledge
Interpretation	Design based on domain knowledge, strong interpretative ability	Outputs only detection results, weak interpretative ability
Unknown attack detection	Only detects known attacks	Detects known and unknown attacks

As seen in Figure 4, misuse identification involves pattern-based, specialist framework and finite-state machine-based approaches for taxonomy-based Detection. Anomaly analysis involves approaches focused on mathematical simulation, machine learning, and time series.

3.3 Classification by Source of Data

A benefit of host-based IDSs is that they can reliably detect intrusions and respond to them since they can track substantial object activity (e.g. confidential data, programmes, ports). The drawbacks are that host-based IDSs use server resources, rely on the stability of the host and cannot detect network attacks. Usually, a network-based IDS is used on large hosts or switches. Many network IDSs are independent of the O.S. and can thus be found in various operating systems environments. Besides, network-based IDSs can detect some types of protocols and network attacks. The downside is that they only track the traffic that goes through a particular network path. Table 4 demonstrates the significant variations between host-based IDS and network-based IDS.

© 2020, IJCERT All Rights Reserved https://doi.org/10.22362/ijcert/2020/v7/i12/v7i1203

Table 4 Differences between host-based IDS and network-based IDS

	Host-based IDS	Network-based IDS
Source of data	Logs of the operating system or application programs	Network traffic
Deployment	Every host; Dependent on operating systems; Difficult to deploy	Key network nodes; Easy to deploy
Detection efficiency	Low, must process numerous logs	High, can detect attacks in real-time
Intrusion traceability	Trace the process of intrusion according to system call paths	Trace position and time of intrusion according to I.P. addresses and timestamps
Limitation	Cannot analyze network behaviors	Monitor only the traffic passing through a specific network segment

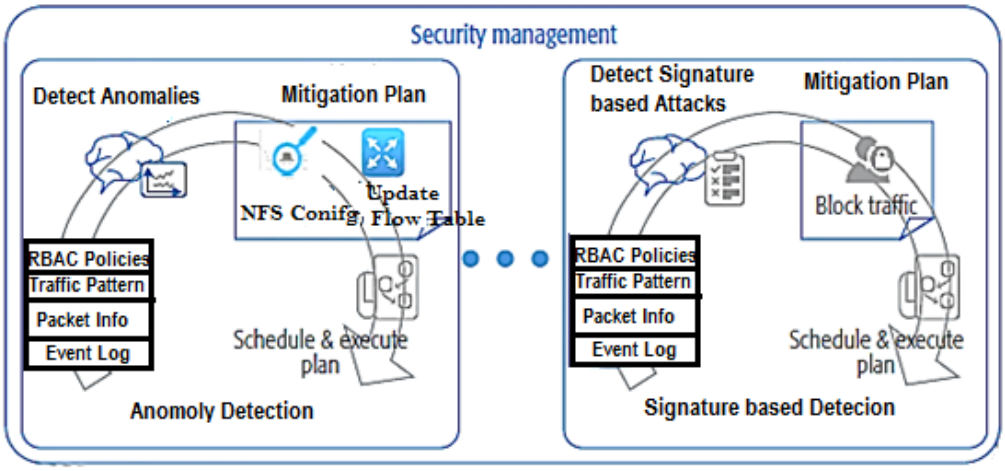

Figure 6. Security management Architecture

Security Management: The most common security method includes scanning the network for trends and emerging attacks. Vulnerability management: This leaves the network vulnerable to zero-day attacks, though. This vulnerability is essential as new threats occur every day. The need for rigorous safety measures is evident, and the position of ML in this regard was thoroughly investigated in Figure 6. Current studies have centred on the use of ML for misuse detection to study complicated attack patterns from historical data and develop basic rules for the identification of differences in documented attacks. ML detection of anomalies to detect zero-day attacks was also discussed. This consists of studying standard behaviour patterns and identifying deviations from the average.

4.Common Machine Learning Algorithms in Ids

4.1 Machine Learning Models

Two significant forms of machine learning exist supervised and unregulated learning. Learning supervised depends on valuable knowledge in classified results. Classification is the most common task in supervised learning (which is used most commonly in IDS); however, manual classification of data is costly and time-consuming. The absence of adequate classified data thus constitutes the key bottleneck for supervised learning. In comparison, unattended research derives useful knowledge from unlabelled data, making training results much more comfortable to access. However, unmonitored learning performance is significantly lower than in

© 2020, IJCERT All Rights Reserved https://doi.org/10.22362/ijcert/2020/v7/i12/v7i1203

supervised learning performance. Figure 7 displays the popular machine learning algorithms used in IDSs.

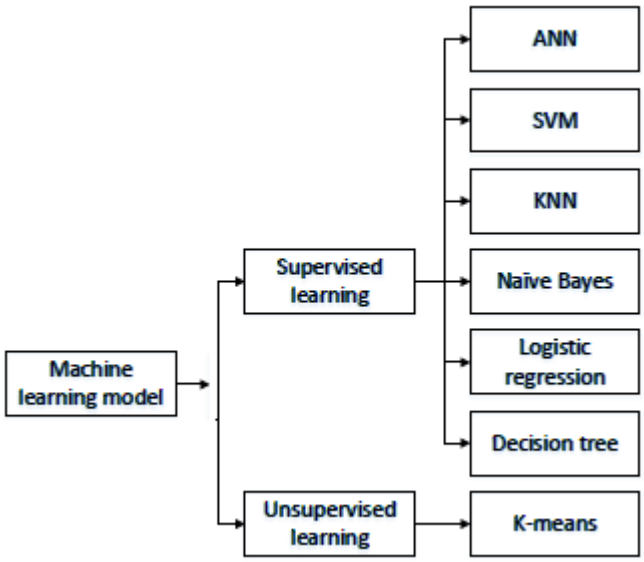

Figure 7. Classification of machine learning algorithms.

Table 5 The advantages and Disadvantages of various ML classification Algorithms

Algorithm	Advantages	Disadvantages	Improvement Measures
ANN	Able to deal with nonlinear data; Strong fitting ability	Suitable to over fitting; Likely to become confined in a local optimum; Model training is time consuming	Adopted improved optimizers, activation functions, and loss functions
SVM	Learn useful information from small train set; Strong generation capability	Do not perform well on big data or multiple classification tasks; Sensitive to kernel function parameters	Optimized parameters by particle swarm optimization (PSO)[25]

© 2020, IJCERT All Rights Reserved https://doi.org/10.22362/ijcert/2020/v7/i12/v7i1203

KNN	Apply to massive data; Suitable to nonlinear data; Train quickly; Robust to noise	Low accuracy on the minority class; Long test times; Sensitive to the parameter K	Reduced comparison times by trigonometric inequality; Optimized parameters by particle swarm optimization (PSO) [26]; Balanced datasets using the synthetic minority oversampling technique (SMOTE) [27]
Naïve Bayes	Robust to noise; Able to learn incrementally	Do not perform well on attribute-related data	Imported latent variables to relax the independent assumption [28]
LR	Simple, can be trained rapidly; Automatically scale features	Do not perform well on nonlinear data; Apt to overfitting	Imported regularization to avoid overfitting [28]
Decision tree	Automatically select features; Strong interpretation	Classification result trends to majority class; Ignore the correlation of data	Balanced datasets with SMOTE; Introduced latent variables
K-means	Simple, can be trained rapidly; Strong scalability; Can fit to big data	Do not perform well on nonconvex data; Sensitive to initialization; Sensitive to the parameter K	Improved initialization method [30]

4.2. Performance Evaluation Metrics (PEM)

Several criteria are used for evaluating methods of machine learning. These metrics are used to choose the best models. Multiple metrics are also concurrently used in IDS research to accurately quantify the identification effect.

Accuracy is defined as the ratio of samples to total samples correctly identified. Accuracy is an effective criterion for managing dataset. However, in actual network settings, normal samples are far more frequent than irregular samples; precision might not be the acceptance criterion.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

Precision (P) is defined as the relationship between true positive samples and forecast positive samples; it constitutes confidence in the Detection of attacks.

$$P = \frac{TP}{TP+FP}$$

© 2020, IJCERT All Rights Reserved https://doi.org/10.22362/ijcert/2020/v7/i12/v7i1203

Recall (R) is defined as the ratio of true positive samples to total positive samples, and the detection rate is often called. The detection rate represents the ability of the model to detect attacks, an essential parameter in IDS.

$$R = \frac{TP}{TP+FN}$$

F-measure (F) is defined as the harmonic average of the precision and the recall.

$$F = \frac{2*P*R}{P+R}$$

The false-negative rate (FNR) The FNR is defined as the ratio of False Negative Samples to Total Positive Samples. The FNR is also regarded as the missing warning rate in threat detection.

$$FNR = \frac{FN}{TP + FN}$$

The false-positive ratio (FPR) is defined as the ratio of false-positive to positive samples expected. The FPR is also called the false alarm rate in attack detection, and it is measured as follows:

$$FPR = \frac{FP}{TP + FP}$$

Where the T.P. true positives, F.P. are the false positives, T.N. is the true negatives, and F.N. is the false negatives. An IDS seeks to distinguish attacks; thus, samples of attacks are typically considered positive and standard samples are typically considered negatives. The methods used in attack detection include accuracy, recall (or identification rate), FNR (or missing alarm rate), and FPR (or incorrect alarm rate).

4.3 Benchmark Datasets in IDS

The goal of machine learning is to derive useful knowledge from data; thus, machine learning success depends on the consistency of the input data. Data comprehension is the foundation of the technique for machine learning. For IDSs, the data obtained should be easy to obtain and show hosts/networks behaviour. Packets flow, sessions, and logs are typical data sources for IDSs. The design of the data collection is complicated and time-consuming. If a benchmark dataset is created, many researchers may repeatedly reuse it. In addition to simplicity, the use of comparison data sets has two other advantages.

(1) Authoritative comparison data sets make experimental findings more compelling.

(2) Several recent studies have been carried out using standard comparison datasets that make it easy to compare current findings with previous research.

(1) DARPA1998

The DARPA1998 dataset [31] was developed by the MIT Lincoln Laboratory and is a commonly used benchmark dataset in IDS studies. Five labels are available: standard, denial of service (DOS), sample, User to Root (U2R) and local remote (R2L). Since raw packets cannot be added directly to standard machine learning models, this limitation has been solved by the KDD99 dataset.

(2) KDD99

The KDD99 [32] dataset is the most commonly used IDS benchmark. In KDD99, the codes are similar to DARPA1998. In KDD99, there are four kinds of features, i.e. essential characteristics, application attributes, host-based statistics, and time-based statistics. Unfortunately, there are several flaws in the KDD99 dataset. KDD data is too old for the new network environment to reflect.

(3) NSL-KDD

The NSL-KDD[38] was introduced to fix the limitations of the KDD99 data collection. The NSL-KDD documents have been carefully chosen based on KDD99. Different class records in the NSL-KDD are matched, which eliminates the classification bias problem. The NSL-KDD has excluded duplicate and obsolete documents, which ensures that the number of records is only small. The NSL-KDD alleviates, to some extent, the issues of data fragmentation and redundancy. However, new data is not included in the NSL-KDD; thus, samples of minority groups appear to lag, and their samples are obsolete.

(4) ISCX 2012

Dedicated network monitoring has been tested for regular device activity on TCP, SMTP, SSH, IMAP, POP3 and FTP protocols in this dataset [52] (Shiravi et al., 2012). This dataset is based on realistic, classified network traffic involving different possibilities for attacks.

(5) UNSW-NB15

The University of South Wales has assembled the UNSW-NB15 [33] dataset, where researchers set up three virtual servers to collect network traffic and extracted 49-dimensional features using the Bro platform. The dataset incorporates more types of attacks than the KDD99 dataset and has more features. Although the impact of UNSW-NB15 is currently less than that of KDD99, new data sets are required for the creation of new machine learning IDS.

(6) CICIDS 2017

CICIDS2017's dataset contains all brain and new Malware threat information including Brute Force FTP, Brute Force SSH, DoS, Heart bleed, Network Assault, Invasion, Botnet and DDoS [53]. This dataset is categorized according to timestamps, I.P.s, source and destination ports, protocols and attacks, respectively.

5. Research on Machine Learning-Based IDs

© 2020, IJCERT All Rights Reserved https://doi.org/10.22362/ijcert/2020/v7/i12/v7i1203

Machine learning is a type of data-driven approach, where the first step is to understand the data. Therefore we use the kind of data source as the key classification thread and present many ways of applying machine learning to IDS architecture for different kinds of data in this section. Related data types represent multiple attack behaviours, including host and network behaviour. Device logs reflect host habits, and network activities reflect network activity. There are several types of attacks, each with a particular pattern. Therefore, it is essential to choose suitable data sources to identify various attacks according to the threat characteristics.

The primary function of a DOS attack, for example, is to transmit many packets over a short time; hence flow data is ideal for the identification of a DOS attack. A secret channel includes data-release operations between two separate I.P. addresses that are more appropriate for session data detection.

5.1 Packet-Based Attack Detection

Packets, which are the basic units of network communication, represent each communication's information. Packets consist of binary files, which means they are not readable until first scanned. A packet consists of a programme header and data. The headers are standardized fields that define I.P., ports and other protocol-specific fields. The application data section requires payload protocols from the application layer.

There are three benefits of using packets as IDS data sources: (1) packets contain contact information, so U2L and R2L attacks can be easily detected. (2) Packets contain I.P.s and timestamps so that attack origins can be identified correctly. (3) Packs can be read without caching automatically and can thus be tracked in real-time. However, each packet does not represent the complete state of communication or the context of each packet, so it is challenging to identify attacks like DDOS. Packet-based identification techniques primarily provide packet decoding and payload processing approaches.

5.2 Machine learning Centric Secure Cloud Management

Table 4 Machine learning Centric Secure Cloud Management

Cloud Management Area (CMA)	Cloud Management function (CMF)	ML Techniques
Security	Signature-based Detection	NN,DT,BN,SVM
	Anomaly Detection	(Collaborative) NN,DNN,k-NN,K-means, (Collaborative) DT, (Collaborative) BN, SVM

© 2020, IJCERT All Rights Reserved

Table 4. Describes the Centric Protected Cloud Security Machine Learning for two forms of IDS classification, i.e. signatures and anomaly-based threat identification. Techniques such as N.N., D.T., B.N., SVM, N.N., DNN, K-means, (Collaborative) D.T., (Collaborative) B.N., and SVM for the identification of an abnormality are proposed.

Table 5 Summary of the machine learning-based IDSs Classification Methods.

Authors	Data Source	Datasets	Classification Methods	Machine Learning Algorithms
Mayhew et al. [34]	Packet	Private dataset	Packet parsing	SVM and K-means
Hu et al. [35]	Packet	DARPA 2000	Packet parsing	Fuzzy C-means
Min et al. [36]	Packet	ISCX 2012	Payload analysis	CNN
Zeng et al. [37]	Packet	ISCX 2012	Payload analysis	CNN, LSTM, and auto-encoder
Yu et al. [38]	Packet	CTU-UNB	Payload analysis	Auto-encoder
Rigak et al. [39]	Packet	Private dataset	Payload analysis	GAN
Goeschel et al. [40]	Flow	KDD99	Statistic feature for flow	SVM, decision tree, and Naïve Bayes
Kuttranont et al. [41]	Flow	KDD99	Statistic feature for flow	KNN
Peng et al. [42]	Flow	KDD99	Statistic feature for flow	K-means
Teng et al. [43]	Flow	KDD99	Traffic grouping	SVM
Ma et al. [44]	Flow	KDD99 and NSL-KDD	Traffic grouping	DNN
Ahmim et al. [45]	Session	CICIDS 2017	Statistic feature for session	DT
Alseiari et al. [46]	Session	Private dataset	Statistic feature for session	K-means
Yuan et al. [47]	Session	ISCX 2012	Sequence feature for session	CNN and LSTM
Radford et al. [48]	Session	ISCX IDS	Sequence feature for session	LSTM
Wang et al. [49]	Session	DARPA 1998 and ISCX 2012	Sequence feature for session	CNN
Meng et al. [50]	Log	Private dataset	Rule-based	KNN
McElwee et al. [51]	Log	Private dataset	Rule-based	DNN

6. Cloud Security Challenges

The complexities of cloud protection are part of ongoing research. Related open concerns as potential ranges are identified:

Security-based Data Classification: A cloud storage data centre can contain data from multiple users. Classification of data may be achieved to include the protection standard depending on the value of data. This classification system should discuss multiple issues such as frequency of access, frequency of changes and access for separate organizations based on a data type. The protection level associated with this tagged data feature can be added after the data has been identified and tagged. The protection standard requires authentication, encryption, privacy and storage etc. chosen depending on the data form.

Identity management system: A secure, trust-based identity management system is essential for both providers and consumers of cloud services. Related identity management system problems have been reported. The approach to ensure the identification and delivery, preservation and control of life cycles is a call for a trust-based identity management scheme.

Safe, cloud infrastructure solution: a stable cloud storage system, combined with overall security concerns, is a challenge. A stable and trustworthy approach is the prerequisite that the cloud computing system has to solve.

Technology Optimization Uses: Security issues and virtualization requirements must also be discussed and tackled along with efficient usage of the cloud resources.

7. Conclusion

In this paper, we investigate emerging cloud protection challenges and new security solutions. We recognize 28 cloud protection problems such as firewalls, malicious insiders, faulty plugins, multi-tenancy applications, side channels, insecure browsers, and

© 2020, IJCERT All Rights Reserved

usability. We then categorize these concerns into five categories of protection, including security requirements, network, connectivity, cloud infrastructure and data.

The data and cloud resources should be secure from known / unknown attacks in all cloud components to achieve robust cloud protection. More analysis is being conducted to address security issues in the cloud world. However, there are also many open challenges to be addressed to have a stable cloud infrastructure. Security considerations relating to cloud networking, network, anonymity, and application and web resources are some of the conventional challenges at the onset of cloud computing.

Machine learning models play an enormously important role and have been an excellent research path. Also, we explore the IDS taxonomy, which uses data sources as a central thread to present the various algorithms used in this field for machine learning. We then expand and address IDSs extended to different data streams, i.e. logs, packets, flow, and sessions, based on these taxonomies. To detect threats, IDSs must then choose the right source of data according to the threat characteristics.

References

[1] https://www.prnewswire.com/news-releases/the-global-cloud-security-market-to-reach-usd-1264-billion-by-2024-300558185.html (Accessed on 10th April 2020)

[2] Subramanian N, Jeyaraj A (2018) Recent security challenges in cloud computing. Compute Electr Eng 71:28–42

[3] Mell P, Grance T (2018) SP 800-145, The NIST Definition of cloud computing | CSRC (online) Csrc.nist.gov. https://csrc.nist.gov/publications/detail/sp/800-145/fnal. Accessed 11 Dec 2018

[4] Xu X (2012) From cloud computing to cloud manufacturing. Robot Comput Integr Manuf 28(1):75–86.

[5] Bhamare D, Samaka M, Erbad A, Jain R, Gupta L, Chan HA (2017) Optimal virtual network function placement in multi-cloud service function chaining architecture. Comput Commun 102:1–16

[6] Michie, D.; Spiegelhalter, D.J.; Taylor, C.(1994) Machine Learning, Neurall and Statistical Classification; Ellis Horwood Series in Artificial Intelligence: New York, NY, USA, Volume 13.

[7] Buczak, A.L.; Guven, E.(2015) A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Commun. Surv. Tutor. 18, 1153–1176.

[8] Xin, Y.; Kong, L.; Liu, Z.; Chen, Y.; Li, Y.; Zhu, H.; Gao, M.; Hou, H.;Wang, C.(2018) Machine learning and deep learning methods for cybersecurity. IEEE Access, 6, 35365–35381.

[9] Agrawal, S.; Agrawal, J.(2015) Survey on anomaly detection using data mining techniques. Procedia Comput. Sci., 60, 708–713.

[10] Sengupta, S.; Kaulgud, V.; Sharma, V.S.(2011) Cloud computing security Trends and research directions. In Proceedings of the IEEE World Congress on Services (SERVICES), Washington, DC, USA, 4–9; pp. 524–531.

[11] Tripathi, A.; Mishra, A(2011) Cloud computing security considerations. In Proceedings of the IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Xi'an, China, 14–16 , pp. 1–5.

[12] Morin, J.; Aubert, J.; Gateau, B. (2012) "Towards cloud computing SLA risk management: Issues and challenges". In Proceedings of the 45th Hawaii International Conference on System Science (HICSS), Maui, HI, USA, 4–7; pp. 5509–5514.

[13] Braun, V.; Clarke, V. (2006) Using thematic analysis in psychology. *Qual. Res. Psychol.* , 77–101.

[14] A Survey on Cloud Computing Security, Challenges and threats|Whitepapers|TechRepublic. Available online: http://www.techrepublic.com/whitepapers/a-survey-on-cloud- computingsecurity- challenges-and-threats/3483757 (accessed on 18 April 2020).

[15] Thalmann, S.; Bachlechner, D.; Demetz, L.; Maier, R.(2012) "Challenges in cross-organizational security management". In Proceedings of the 45th Hawaii International Conference on System Science (HICSS), Maui, HI, USA, 4–7; pp. 5480–5489.

[16] Wang, J.-J.; Mu, S.(2011) Security issues and countermeasures in cloud computing. In Proceedings of the IEEE International Conference on Grey Systems and Intelligent Services (GSIS), Nanjing, China, 15–18 ; pp. 843–846.

[17] Lv, H.; Hu, Y.(2011) "Analysis and research about cloud computing security protect policy". In Proceedings of the International Conference on Intelligence Science and Information Engineering (ISIE), Wuhan, China, 20–21; pp. 214–216.

[18] Jain, P.; Rane, D.; Patidar, S.(2011) A survey and analysis of cloud model-based security for computing secure cloud bursting and aggregation in renal

© 2020, IJCERT All Rights Reserved

environment. In Proceedings of the World Congress on Information and Communication Technologies (WICT), Mumbai, India, 11– 14; pp. 456–461.

[19] Behl, A.(2011) Emerging security challenges in cloud computing: An insight to cloud security challenges and their mitigation. In Proceedings of the 2011 World Congress on Information and Communication Technologies (WICT), Mumbai, India, 11–14; pp. 217–222.

[20] Mathisen, E.(2011) Security challenges and solutions in cloud computing. In Proceedings of the 5th IEEE International Conference on Digital Ecosystems and Technologies Conference (DEST), Daejeon, Korea; pp. 208–212.

[21] Mahmood, Z. (2011) Data location and security issues in cloud computing. In Proceedings of the International Conference on Emerging Intelligent Data and Web Technologies (EIDWT), Tirana,Albania, 7–9; pp. 49–54.

[22] Denning, D.E(1987) An intrusion-detection model. IEEE Trans. Softw. Eng. 222–232.

[23] Heberlein, L.T.; Dias, G.V.; Levitt, K.N.; Mukherjee, B.; Wood, J.; Wolber, D.(1990) A network security monitor. In Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy, Oakland, CA, USA, 7–9; pp. 296–304.

[24] Kuang, F.; Zhang, S.; Jin, Z.; Xu,W.(2015) A novel SVM by combining kernel principal component analysis and improved chaotic particle swarm optimization for intrusion detection. Soft Comput., 19, 1187–1199.

[25] Syarif, A.R.; Gata, W.(2017) Intrusion detection system using hybrid binary PSO and K-nearest neighborhood algorithm. In Proceedings of the 2017 11th International Conference on Information & Communication Technology and System (ICTS), Surabaya, Indonesia,; pp. 181–186.

[26] Pajouh, H.H.; Dastghaibyfard, G.; Hashemi, S.(2017) Two-tier network anomaly detection model: A machine learning approach. J. Intell. Inf. Syst. 48, 61–74.

[27] Mahmood, H.A.(2018) Network Intrusion Detection System (NIDS) in Cloud Environment based on Hidden Naïve Bayes Multiclass Classifier. Al-Mustansiriyah J. Sci., 28, 134–142.

[28] Shah, R.; Qian, Y.; Kumar, D.; Ali, M.; Alvi, M.(2017) Network intrusion detection through discriminative feature selection by using sparse logistic regression. Future Internet, 9, 81.

[29] Peng, K.; Leung, V.C.; Huang, Q.(2018) Clustering approach based on mini batch kmeans for intrusion detection system over big data. IEEE Access, 6, 11897–11906.

[30] DARPA1998 Dataset. 1998. Available online: http://www.ll.mit.edu/r-d/datasets/1998-darpa-intrusiondetection- evaluation-dataset (accessed on 16 March 2020).

[31] KDD99 Dataset. 1999. Available online: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (accessed on 16 March 2020).

[32] NSL-KDD99 Dataset. 2009. Available online: https://www.unb.ca/cic/datasets/nsl.html (accessed on 16 March 2020).

[33] Mayhew, M.; Atighetchi, M.; Adler, A.; Greenstadt, R.(2015) Use of machine learning in big data analytics for insider threat detection. In Proceedings of the MILCOM 2015-2015 IEEE Military Communications Conference, Canberra, Australia; pp. 915–922.

[34] Hu, L.; Li, T.; Xie, N.; Hu, J. (2015) False positive elimination in intrusion detection based on clustering. In Proceedings of the 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Zhangjiajie, China; pp. 519–523.

[35] Min, E.; Long, J.; Liu, Q.; Cui, J.; Chen, W.(2018), TR-IDS: Anomaly-based intrusion detection through text-convolutional neural network and random forest. Secur. Commun. Netw, 4943509.

[36] Zeng, Y.; Gu, H.; Wei, W.; Guo, Y. Deep (2019) Full Range: A Deep Learning Based Network Encrypted Traffic Classification and Intrusion Detection Framework. IEEE Access, 7, 45182–45190.

[37] Yu, Y.; Long, J.; Cai, Z.(2017) Network intrusion detection through stacking dilated convolutional autoencoders. Secur. Commun. Netw. **2017**, 2017, 4184196.

[38] Rigaki, M.; Garcia, S.(2018) Bringing a gan to a knife-fight: Adapting malware communication to avoid Detection. In Proceedings of the 2018 IEEE Security and PrivacyWorkshops (SPW), San Francisco, CA, USA, pp. 70–75.

[39] Goeschel, K.(2016) Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and naive Bayes for off-line analysis. In Proceedings of the SoutheastCon 2016, Norfolk, VA, USA,; pp. 1–6.

© 2020, IJCERT All Rights Reserved

[40] Kuttranont, P.; Boonprakob, K.; Phaudphut, C.; Permpol, S.; Aimtongkhamand, P.; KoKaew, U.;Waikham, B.; So-In, C.(2017) Parallel KNN and Neighborhood Classification Implementations on GPU for Network Intrusion Detection. J. Telecommun. Electron. Comput. Eng. (JTEC), 9, 29–33.

[41] Peng, K.; Leung, V.C.; Huang, Q(2018). Clustering approach based on mini batch kmeans for intrusion detection system over big data. IEEE Access **2018**, 6, 11897–11906.

[42] Teng, S.; Wu, N.; Zhu, H.; Teng, L.; Zhang, W.(2017) SVM-DT-based adaptive and collaborative intrusion detection. IEEE/CAA J. Autom. Sin., 5, 108–118.

[43] Ma, T.; Wang, F.; Cheng, J.; Yu, Y.; Chen, X(2016) A hybrid spectral clustering and deep neural network ensemble algorithm for intrusion detection in sensor networks. Sensors **2016**, 16, 1701.

[44] Ahmim, A.; Maglaras, L.; Ferrag, M.A.; Derdour, M.; Janicke, H.(2019) A novel hierarchical intrusion detection system based on decision tree and rules-based models. In Proceedings of the 2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS), Santorini Island, Greece, pp. 228–233.

[45]. Alseiari, F.A.A.; Aung, Z. (2015) Real-time anomaly-based distributed intrusion detection systems for advanced Metering Infrastructure utilizing stream data mining. In Proceedings of the 2015 International Conference on Smart Grid and Clean Energy Technologies (ICSGCE), Offenburg, Germany,pp. 148–153.

[46]. Yuan, X.; Li, C.; Li, X.(2017) DeepDefense: identifying DDoS attack via deep learning. In Proceedings of the 2017 IEEE International Conference on Smart Computing (SMARTCOMP), Hong Kong, China; pp. 1–8.

[47] Radford, B.J.; Apolonio, L.M.; Trias, A.J.; Simpson, J.A.(2018) Network traffic anomaly detection using recurrent neural networks. arXiv:1803.10769.

[48] Wang, W.; Sheng, Y.; Wang, J.; Zeng, X.; Ye, X.; Huang, Y.; Zhu, M.(2017) HAST-IDS: Learning hierarchical spatial-temporal features using deep neural networks to improve intrusion detection. IEEE Access , 6, 1792–1806.

[49] Meng, W.; Li, W.; Kwok, L.F(2015) .Design of intelligent KNN-based alarm filter using knowledge-based alert verification in intrusion detection. Secur. Commun. Netw. 8, 3883–3895.

[50] McElwee, S.; Heaton, J.; Fraley, J.; Cannady, J.(2017) Deep learning for prioritizing and responding to intrusion detection alerts. In Proceedings of the MILCOM 2017—2017 IEEE Military Communications Conference (MILCOM), Baltimore, MD, USA, pp. 1–5.

[51] Shiravi A, Shiravi H, Tavallaee M, Ghorbani AA (2012) Toward developing a systematic approach to generate benchmark datasets for intrusion detection. Computers & security 31(3):357–374

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

© 2020, IJCERT All Rights Reserved