Data & Metadata. 2022; 1:34
doi: 10.56294/dm202272

REVIEW

Anomaly Detection in Network Traffic using Machine Learning for Early Threat Detection

Detección de anomalías en el tráfico de red mediante aprendizaje automático para la detección temprana de amenazas

Mohammed Hussein Thwaini[1] ✉

[1]University of Fallujah, Applied Sciences College, Iraq. 1 University of Fallujah, Applied Sciences College, Iraq.

Cite as: Thwaini MH. Anomaly Detection in Network Traffic using Machine Learning for Early Threat Detection. Data and Metadata. 2022;1:34. https://doi.org/10.56294/dm202272

Submitted: 17-09-2022 **Revised:** 29-11-2022 **Accepted:** 21-12-2022 **Published:** 23-12-2022

Editor: Prof. Dr. Javier González Argote

ABSTRACT

Due to advances in network technologies, the amount of people using networks is rising rapidly. This has resulted in a large amount of transmission information being generated and moved across the network. However, this data is vulnerable to attacks and intrusions. To prevent network intrusions, security measures must be implemented, which can detect anomalies and identify potential threats. Network security researchers and labs have done extensive research in network security. The purpose of this study was to perform a noninvasive inspection to give a large general mechanism on recent advances in abnormality detection. The study reviewed recent research published in the past five years, which examined new technologies and potential future opportunities in anomaly detection. The literature review focused specifically on anomaly detection systems used in network traffic. This included various applications such as Wireless Sensor Networks (WSN), Internet of Things (IoT), High Performance Computing, Industrial Control Systems (ICS), and Software Defined Networking (SDN) environments. The review concludes by highlighting several unresolved issues that need to be addressed in order to improve anomaly detection systems.

Keywords: Anomaly Detection; Intrusion; Networks; Supervised; Unsupervised.

RESUMEN

Debido a los avances en las tecnologías de red, el número de personas que las utilizan está aumentando rápidamente. Esto ha dado lugar a que se genere una gran cantidad de información de transmisión que se mueve a través de la red. Sin embargo, estos datos son vulnerables a ataques e intrusiones. Para evitar las intrusiones en la red, es necesario aplicar medidas de seguridad que permitan detectar anomalías e identificar posibles amenazas. Los investigadores y laboratorios de seguridad de redes han realizado numerosas investigaciones en este ámbito. El objetivo de este estudio era realizar una inspección no invasiva para ofrecer un amplio mecanismo general sobre los avances recientes en la detección de anomalías. El estudio revisó la investigación reciente publicada en los últimos cinco años, que examinó las nuevas tecnologías y las posibles oportunidades futuras en la detección de anomalías. La revisión bibliográfica se centró específicamente en los sistemas de detección de anomalías utilizados en el tráfico de red. Esto incluía diversas aplicaciones como redes de sensores inalámbricos (WSN), Internet de las cosas (IoT), informática de alto rendimiento, sistemas de control industrial (ICS) y entornos de redes definidas por software (SDN). La revisión concluye destacando varias cuestiones no resueltas que deben abordarse para mejorar los sistemas de detección de anomalías.

Palabras clave: Detección De Anomalías; Intrusión; Redes; Supervisada; No Supervisada.

INTRODUCTION

With the ever-increasing reliance on computer networks for various operations, the importance of network security has become paramount. Organizations across industries face the constant challenge of protecting their networks from potential threats, such as malware infections, unauthorized access, and data breaches. Traditional security measures like firewalls and intrusion detection systems can mainly handle known threats. However, these traditional approaches often fail to identify novel and evolving threats.

To combat this issue, machine learning techniques have gained significant attention in recent years. Machine learning algorithms can analyze large volumes of network traffic data and identify anomalous patterns that may indicate potential security threats. By leveraging historical data and learning from past incidents, these algorithms can identify deviations from normal network behavior, thus enabling early threat detection and proactive action.(1)

The primary objective of using machine learning for network traffic anomaly detection is to develop models that can accurately classify network traffic as either normal or malicious. It involves training models using labeled datasets of network traffic and various features such as packet size, protocol type, source/destination IP addresses, and port numbers. These models can then be deployed in real-time to continuously monitor network traffic and generate alerts whenever anomalous behavior is detected.(2)

The benefits of using machine learning for network traffic anomaly detection are numerous. It allows organizations to detect threats that traditional security measures might miss, providing an additional layer of security. Moreover, machine learning can adapt and learn from new data, improving its accuracy over time to handle evolving threats. It also helps minimize false positives by reducing the number of unnecessary alerts, enabling security teams to focus on genuine threats more effectively.

machine learning techniques have emerged as a powerful tool for detecting network traffic anomalies and enabling early threat detection. By analyzing large volumes of network traffic data and identifying abnormal patterns, these algorithms can enhance network security and mitigate potential threats. Incorporating machine learning into network security strategies can assist organizations in staying one step ahead of attackers and safeguarding their valuable digital assets

Related works

Chandola et al. studied Network anomalies are abnormal occurrences within the network that deviate from normal or known behavior and are believed to have security implications. Also known as atypical actions intended to disrupt the regular operations of a network. Anomalies are identified by patterns in the data that do not conform to a clearly defined concept of logical state.(3)

Zhao et al., It can be said that irrational and significant abnormalities in the transmission mechanism are said to be anomalies.(4)

Ahmad et al., 2017 It is an instantaneous stage in which the actions of the system take place deviates significantly from its previous normal behavior.(5)

Mohd Ali, 2018 Different authors have used various terms such as abnormalities, outliers, or exceptions to refer to network anomalies, leading to confusion in the terminology. To grasp the concept of anomalies in a network system, it is essential to understand what is considered normal. There are three main types of network anomalies: point anomalies, contextual anomalies, and collective anomalies.(6)

METHODS

Ensemble methods, also known as multi-classifier systems, involve training multiple machine learning models to create streamlined parachute ornaments and then a nesting occurs results to improve precision (Aburomman et al., 2017). Several intrusion detection systems (IDSs) have been It was developed on the principle of clustering mechanisms, as noted across the literature.

One such system, developed by Gu et al. (2019) is a good SVM data-driven IDS that includes feature augmentation. They applied intensity ratios to the baseline curves, which improved the quality of the training data. Experimental results showed that the SVM group achieved a competitive performance in terms of the results of the accuracy of observations in monitoring, training trained quickly, clarity, and false news giving data compared to other data. NSL-KDD dataset was used for evaluation.

Pham et al. (2018) Provide data within a specific structure and mechanism for taking advantage aimed at improving IDS performance. They used clustering mechanisms with tree-based mechanisms as the main data. Its operation showed that the fill through the J48 data improved the classification accuracy and reduced the false alarm rate in the NSL-KDD datasets.

In another work by Bhatti et al. (2020), a classification-based algorithm was developed to observe the data of the attack variant. The execution boundary consisted of four main points: data stacking, pre-processing, training procedure, and decision-making.

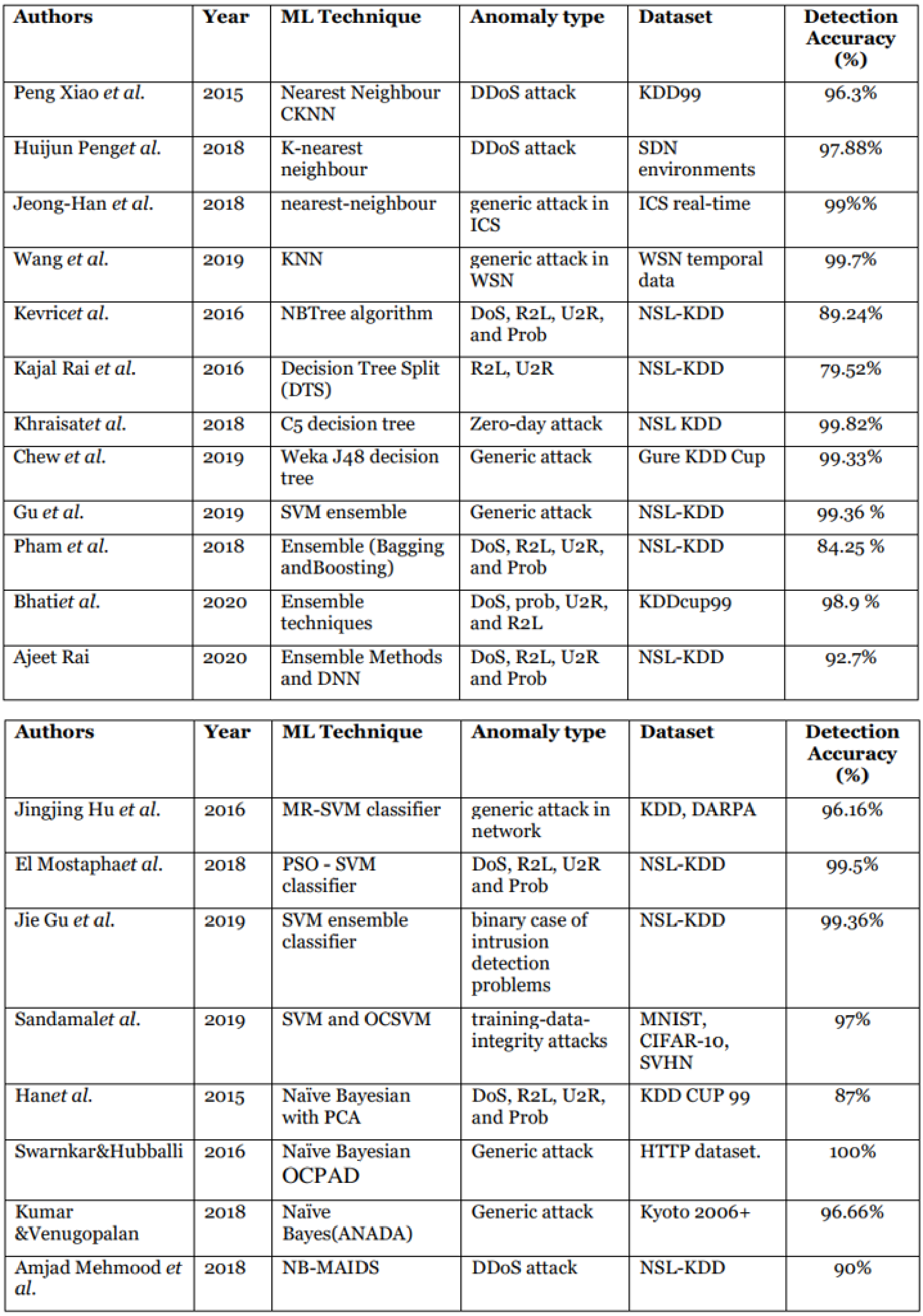
Individual classifiers were trained separately and their decisions were combined using majority voting. The

https://doi.org/10.56294/dm202272

proposed framework achieved high detection accuracy for various attack classes on the KDDcup99 dataset.

Rai (2020) explored ensemble learning methods, such as boosting and bagging algorithms (XGBoost, GBM, and DRF), for IDS. (DNN) was also implemented used the H2O Python library. Genetic algorithm-based feature selection was applied to improve DNN performance. The proposed approach outperformed traditional ML models, and the NSL-KDD dataset was used for evaluation.

Figure 1 provides a comparison of these Research according to the years, the learning mechanism according to the necessary subordination mechanism, the observed abnormal life pattern, the mass of data used.



Authors	**Year**	**ML Technique**	**Anomaly type**	**Dataset**	**Detection Accuracy (%)**
Peng Xiao *et al.*	2015	Nearest Neighbour CKNN	DDoS attack	KDD99	96.3%
Huijun Peng*et al.*	2018	K-nearest neighbour	DDoS attack	SDN environments	97.88%
Jeong-Han *et al.*	2018	nearest-neighbour	generic attack in ICS	ICS real-time	99%%
Wang *et al.*	2019	KNN	generic attack in WSN	WSN temporal data	99.7%
Kevric*et al.*	2016	NBTree algorithm	DoS, R2L, U2R, and Prob	NSL-KDD	89.24%
Kajal Rai *et al.*	2016	Decision Tree Split (DTS)	R2L, U2R	NSL-KDD	79.52%
Khraisat*et al.*	2018	C5 decision tree	Zero-day attack	NSL KDD	99.82%
Chew *et al.*	2019	Weka J48 decision tree	Generic attack	Gure KDD Cup	99.33%
Gu *et al.*	2019	SVM ensemble	Generic attack	NSL-KDD	99.36 %
Pham *et al.*	2018	Ensemble (Bagging andBoosting)	DoS, R2L, U2R, and Prob	NSL-KDD	84.25 %
Bhati*et al.*	2020	Ensemble techniques	DoS, prob, U2R, and R2L	KDDcup99	98.9 %
Ajeet Rai	2020	Ensemble Methods and DNN	DoS, R2L, U2R and Prob	NSL-KDD	92.7%
Authors	**Year**	**ML Technique**	**Anomaly type**	**Dataset**	**Detection Accuracy (%)**
Jingjing Hu *et al.*	2016	MR-SVM classifier	generic attack in network	KDD, DARPA	96.16%
El Mostapha*et al.*	2018	PSO - SVM classifier	DoS, R2L, U2R and Prob	NSL-KDD	99.5%
Jie Gu *et al.*	2019	SVM ensemble classifier	binary case of intrusion detection problems	NSL-KDD	99.36%
Sandamal*et al.*	2019	SVM and OCSVM	training-data-integrity attacks	MNIST, CIFAR-10, SVHN	97%
Han*et al.*	2015	Naïve Bayesian with PCA	DoS, R2L, U2R, and Prob	KDD CUP 99	87%
Swarnkar&Hubballi	2016	Naïve Bayesian OCPAD	Generic attack	HTTP dataset.	100%
Kumar &Venugopalan	2018	Naïve Bayes(ANADA)	Generic attack	Kyoto 2006+	96.66%
Amjad Mehmood *et al.*	2018	NB-MAIDS	DDoS attack	NSL-KDD	90%

Figure 1. Supervised Anomaly detection approaches

Part 1: Network Anomalies Types

Network anomalies are abnormal occurrences within the network that deviate from normal or known behavior and are believed to have security implications. Also known as atypical actions intended to disrupt the regular operations of a network. Anomalies are identified by patterns in the data that do not conform to a clearly defined concept of logical state.[8] It is an instantaneous stage in which the actions of the system take place deviates significantly from its previous normal behavior.[9] It can be said that irrational and significant abnormalities in the transmission mechanism are said to be anomalies.[10]

Different authors have used various terms such as abnormalities, outliers, or exceptions to refer to network anomalies, leading to confusion in the terminology. To grasp the concept of anomalies in a network system, it is essential to understand what is considered normal. There are three main types of network anomalies: point anomalies, contextual anomalies, and collective anomalies.[11]

A point anomaly occurs when a single data point exhibits attributes distinct from the rest of the data group. For instance, If the usual daily spending using a credit card is usually about a hundred dollars, but on a specific day it increases to four hundred dollars, this transaction would be considered an outlier or abnormal occurrence.

"A contextual anomaly "happens when information acts unusually within a particular situation. This kind of anomaly is usually associated with data that changes over time. For instance, if the usual summer enrollment for short courses is between 32 to 45 people, and there are courses with fewer than 15 students, this would be classified as a "contextual anomaly".

A collective anomaly occurs when a cluster of data displays abnormal patterns compared to the rest of the dataset. In this category, abnormalities displayed by individual data points are not taken into account. as an anomaly in itself, but its frequent occurrence within the data is deemed anomalous. For instance, if a computer has a sequence of actions that consistently occur together, such as buffer-overflow, HTTP-web, FTP, HTTP-web, SSH, HTTP-web, SSH, buffer-overflow, HTTP-web, this sequence would be categorized as a collective anomaly.

In conclusion, network anomalies are abnormal network activities that deviate from standard behavior, and they can be classified into point anomalies, contextual anomalies, and collective anomalies. Understanding these types of anomalies is crucial for effectively detecting and mitigating security threats in a network system.

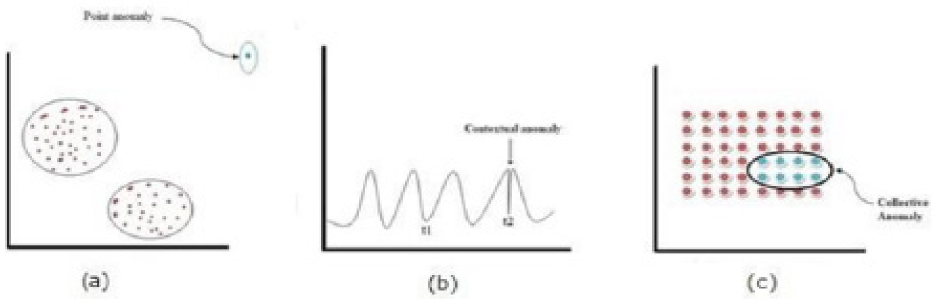


Figure 2. Types of anomalies: "a) point anomaly; b) contextual anomaly;and c) collective anomaly"

Part 2: Anomaly Detection Using Machine Learning

Anomaly detection is a subfield of machine learning that focuses on identifying rare events or patterns in data that deviate significantly from the norm. These anomalies can be indicative of fraudulent activities, system failures, or any unexpected behavior that may require immediate attention.

Machine learning algorithms are trained to recognize patterns and establish a baseline model of what is considered normal behavior within a dataset. Various techniques are then employed to detect deviations from this established norm. The anomalies detected can be either individual data points, groups of data points, or even sequences of events.

There are different types of anomaly detection algorithms, each with its advantages and limitations. One commonly used approach is unsupervised learning, where the algorithm learns patterns from unlabeled data and identifies anomalies based on their deviation from the learned model. This technique is particularly useful when the normal behavior is not well-defined or the dataset is imbalanced.

Another commonly employed method is supervised learning, where the algorithm is trained on labeled data with both normal and anomalous examples. The algorithm learns to differentiate between normal and anomalous patterns based on the labeled information. This technique works well when anomalous instances are well-defined and representative examples are available for training.[12,13]

Apart from these, there are also semi-supervised and ensemble-based approaches that combine the strengths of both unsupervised and supervised techniques. Semi-supervised approaches leverage the labeled examples available along with the unlabeled data to build a model that can detect anomalies effectively. Ensemble-based methods combine multiple models to improve the overall accuracy and robustness of anomaly detection.

Anomaly detection using machine learning finds applications in various domains such as cybersecurity, fraud detection, predictive maintenance, and quality control. In cybersecurity, it can help identify unusual network traffic or unauthorized access attempts. In fraud detection, it can flag fraudulent transactions or activities. In predictive maintenance, it can predict equipment failures by detecting anomalies in sensor data. In quality control, it can identify defective products or anomalies in production processes.(14)

However, it is important to note that anomaly detection using machine learning is a challenging task. Handling high-dimensional data, imbalanced datasets, and evolving patterns can pose significant challenges. Therefore, careful selection of appropriate algorithms, preprocessing techniques, and evaluation metrics is essential to ensure accurate and effective anomaly detection. Continuous monitoring and adaptation of the models are also required to keep up with changing trends and patterns in the data.(15)

Figure 3 illustrates some well-known examples of classification and clustering algorithms.

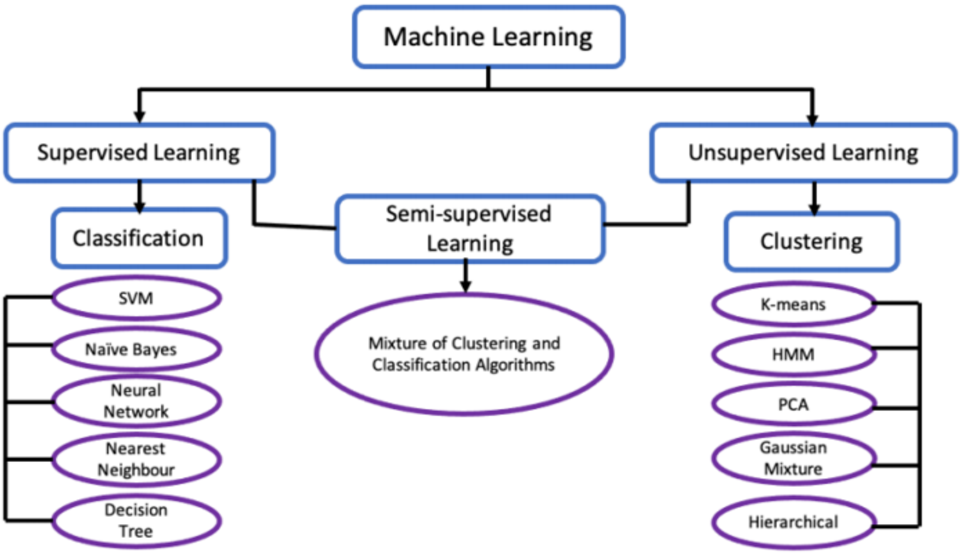
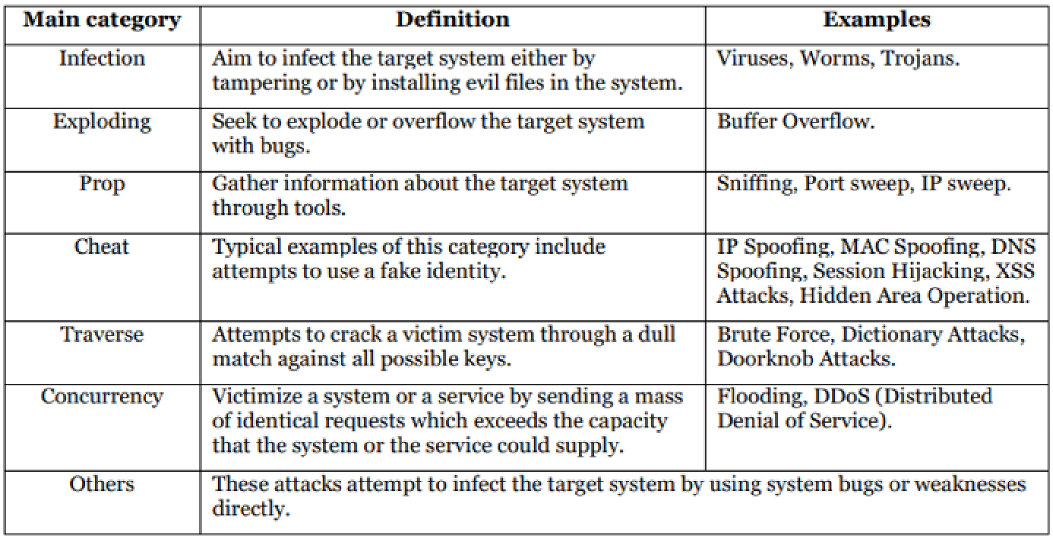


Figure 3. Examples of classification and clustering algorithms

Part 3: Network Attacks

Refers to an illegal attempt to exploit the vulnerability of a computer or network in order to breach the security measures in place,(16) divides attackers into two groups: external and internal. External attackers are unauthorized individuals who target systems, while internal attackers have legitimate access to the system but lack root or superuser privileges,(17) classify attacks into seven main types based on their implementation, as shown in figure 4.

In this survey, our focus will be on the most critical and recent attacks from various categories, along with providing specific examples. We will also explore machine learning (ML) approaches and algorithms used for detecting such attacks. It is important to note that the references cited throughout this text will remain unchanged.



Main category	Definition	Examples
Infection	Aim to infect the target system either by tampering or by installing evil files in the system.	Viruses, Worms, Trojans.
Exploding	Seek to explode or overflow the target system with bugs.	Buffer Overflow.
Prop	Gather information about the target system through tools.	Sniffing, Port sweep, IP sweep.
Cheat	Typical examples of this category include attempts to use a fake identity.	IP Spoofing, MAC Spoofing, DNS Spoofing, Session Hijacking, XSS Attacks, Hidden Area Operation.
Traverse	Attempts to crack a victim system through a dull match against all possible keys.	Brute Force, Dictionary Attacks, Doorknob Attacks.
Concurrency	Victimize a system or a service by sending a mass of identical requests which exceeds the capacity that the system or the service could supply.	Flooding, DDoS (Distributed Denial of Service).
Others	These attacks attempt to infect the target system by using system bugs or weaknesses directly.	

Figure 4. Attack categories

Refers to the use of algorithms in training models with labeled data in domain gridded datasets. These techniques are essential for detecting anomalies and intrusions in network traffic. Several efficient and efficient supervised algorithms are used for this purpose, including Support Vector Machine (SVM), Artificial Neural Network (ANN), Nearest Neighbor Algorithm, Decision Trees, Nearest Neighbors, Cluster Classifiers and Naïve Bayes Classifier. These algorithms are commonly used to detect anomalies using a supervised learning approach. In this review, we provide a summary of recent research in the past five years that have used these supervised learning algorithms to detect anomalies, while ensuring that references are kept in the same format.

Part 4: Support Vector Machine (SVM)

Support Vector Machine (SVM) is a machine learning algorithm that is used for classification and regression tasks. SVM is a type of supervised learning method where we train a model using labeled data to classify new unseen data into different categories.(18)

The main idea behind SVM is to find an optimal hyperplane that separates the different classes in the input data. The hyperplane with the maximum margin between the classes is considered to be the optimal solution. This hyperplane is obtained by maximizing the distance between the closest data points from different classes, which are known as support vectors. (19)

In the case of linearly separable data, SVM finds a linear hyperplane that separates the classes perfectly. However, in real-world scenarios, data is often not linearly separable. To handle such cases, SVM uses a technique known as the kernel trick. This technique transforms the input data into higher dimensions, making it easier to find a hyperplane that separates the classes. The kernel functions used in SVM include linear, polynomial, sigmoid, and radial basis function (RBF), among others. SVM has several advantages. It is effective in high-dimensional spaces, making it suitable for problems with a large number of features. SVM also provides a good generalization capability, meaning it can perform well on new unseen data. Additionally, SVM can handle noisy data by allowing a certain degree of misclassification.(20)

However, SVM has some limitations too. It can be computationally expensive, especially when dealing with large datasets. SVM also requires careful selection of the kernel function and tuning of its parameters, which can be a challenge. Another limitation is that SVM outputs only the class label and does not provide the probability estimates, unlike some other classification algorithms.

In conclusion, SVM is a powerful machine learning algorithm for classifying and regressing data. It can handle both linearly separable and non-linearly separable data by using the kernel trick. SVM has its own advantages and limitations, and its performance depends on the choice of kernel function and tuning of parameters.

Overall, the aforementioned studies present various approaches leveraging the SVM algorithm for intrusion detection. They demonstrate the potential of combining feature selection techniques, parameter optimization algorithms, and defense mechanisms to enhance the effectiveness and efficiency of SVM classifiers in detecting intrusions in network systems.

Part 5: Naïve Bayes

Worked on Naïve Bayesian (NB) in order to find out the network break through (PCA). The model made use of NB with PCA to take in unknown features. Something very good was achieved by the old NB algorithms, which do not take feature weights into account. The values used were "KDD CUP 99", which included attacks such as "DoS, U2L, R2L and Probe" and the results of the experiment gave an increase in finds with a balanced Naïve Bayes rating.(21)

demonstrated a "Naïve Bayesian" of one class called "OCPAD" for load-based anomaly finding. "OCPAD" is an asset-based process that provides the framework and boundaries of the network in which the undefined payload asset is involved. Numerous experiments have been done, demonstrating that OCPAD performs exceptionally well in detecting anomalies with an increased detection rate and a reasonable false positive rate.(22)

introduced a new Naïve Bayes-based algorithm for detecting attacks in training data. The algorithm was subjected to many data over "Kyoto 2006+" data.(23)

The training dataset consists of thousands of average records and thousands of attacks. The Naïve Bayes model was applied in all tests, allowing access to many knowledge processes and detecting anomalies and deceptions.

by presenting a different approach and by implementing the “NB” algorithm for Breach Knowledge and Identification System, which specifically targets distributed denial-of-service (DDoS) attacks in the Internet of Things (IoT) infrastructure. The proposed approach used a hybrid IDS system "NB-MAIDS" based on the use of NB with a multi-server mechanism and data were taken from sensors to report abnormal node activities. The experiment on the NSL-KDD dataset demonstrated the efficiency of the NB classifier with multiple factors, resulting in better performance in blocking attacks with a lower implementation cost.(24)

Part 6: Nearest Neighbor

Moving on to the Nearest Neighbor classifier, foot. A knowledge mechanism based on CKNN to detect DDoS attacks. The CKNN classifier utilized training data correlation information, resulting in reduced training data size and Effective results in catching DDoS vulnerabilities with minimal response time. The authors tested their method on three types of datasets: broad, real, and KDD99.(25)

Peng et al. (2018) Provides a mechanism for detecting bad traffic based on SDN technique for detecting DDoS anomalies in SDN platforms. The K-nearest neighbor algorithm, in combination with P-value, was applied to detect flows. The experiment demonstrated that the" DPTCM-KNN algorithm" Increasing the monitoring accuracy values based on monitoring the negative flow and reducing the insincere positive rate, making it highly effective in SDN platforms.(26)

Yun et al. (2018) formed a statistical mathematical structure using Nearest-Neighbor Search (NNS) for detecting abnormal activities in Industrial Control System (ICS) networks. The proposed model accurately identified Forms transitions, normal and abnormal, even with slight traffic variations and a minimal false rate. The NNS algorithm worked quickly, making it suitable Monitor over normal normal time in any ICS network pattern.(27)

Shifting to wireless sensor networks (WSNs), Wang et al. (2019) systems method for detecting Improper output on WSN environments using the K-nearest neighbor (KNN) algorithm. The proximity of distance was used to analyze the data and detect Spam data at WSN. They researched the different types of uses and types of attacks in WSNs and used the Qual Net network simulation tool for analysis. The results It was shown that the KNN data reached reasonable detection numbers and relatively low error rate, with the compact proximity mechanism that reduces the size of the data set.(28)

In conclusion, these studies demonstrate the effectiveness of Naïve Bayesian and Nearest Neighbor classifiers in detecting anomalies in network intrusion, payload-based anomalies, DDoS attacks, industrial control systems, and wireless sensor networks.

Part 7: Decision Tree

Decision Trees are widely acknowledged as one of the most commonly employed classification techniques in data mining. In a study conducted by Khraisat et al. (2018), they proposed a data mining technique to minimize the rate of false results in a system. The researchers utilized a C5 decision tree classifier and compared it with other data mining algorithms such as SVM, Naïve Bayes, C4.5, and C5. Their aim was to demonstrate that the C5 algorithm produces the best outcomes in detecting abnormal activities. The experimental results displayed that the C5 decision tree effectively decreased both the rate of false positive and false negative results, leading to improved intrusion detection with high accuracy. The NSL-KDD dataset was employed for the experiments.(29)

In a similar manner, Kevric et al. (2017) developed a combined classifier based on the decision tree algorithm for Intrusion Detection Systems (IDS). They utilized a new version of the KDDCUP'99 dataset called NSL-KDD, and a detection algorithm was used to classify network traffic as normal or abnormal based on 41 characteristics describing network traffic patterns. The authors reported achieving exceptional accuracy in the detection rate by combining Naïve Bayes Tree (NB Tree) with random tree classifiers, using a sum rule scheme. This combined approach outperformed the individual random tree algorithm.(30)

Rai et al. (2016) focused on enhancing the decision tree classifier by addressing feature selection and determining split values. These issues are critical in constructing an efficient classifier. The authors proposed the Decision Tree Split (DTS) algorithm based on the C4.5 classifier to tackle these challenges. The algorithm introduces a new method for selecting split values and offers improved efficiency for signature-based intrusion detection, enabling rapid detection of attacks with minimal feature usage and time cost for model building. Comparative analysis with other algorithms in the literature demonstrated the efficacy of the DTS algorithm for constructing decision trees used in intrusion detection. The experiments were conducted on the NSL-KDD dataset.(31)

In a different approach, Chew et al. (2020) proposed a decision tree classifier based on a sensitive pruning model to address the issue of visibility of tree rules in Network-based Intrusion Detection Systems (NIDS). They modified the pruning algorithm based on the C4.8 decision tree and utilized the Weka J48 decision tree pruning framework. The proposed approach was tested on six versions of the Gure KDD Cup IDS datasets. The evaluation and results revealed two advantages of using the C4.8 decision tree: the ability to maintain privacy in the decision tree by selectively hiding sensitive rules, and the flexibility to handle small changes in the pruning process without affecting feature selection.(32)

Part 8: Neural Network

It is referred to as an artificial type of network its symbol (ANN). is a computational model inspired by the human brain. It consists of interconnected nodes, or neurons, that operate in parallel. These neurons are typically interconnected in complex ways and employ as motivating action (Akhi, 2019; Agrawal et al., 2015).

"Neural Networks" (NNs) can be leveraged for supervised or unsupervised learning. And our whole interest will be within the framework of "supervised learning".[33]

Hodo et al. (2016) proposed (ANN) Threat analysis system for IoT networks. By training ANNs Working across the effects of internet packages, the system can Monitor and stop DDoS attacks. This mechanism achieved good numbers in its results, which included the correct and unrealistic correct rates, effectively identifying various types of attacks.[34]

Veselý & Brechlerová (2009) support the idea that ANNs are suitable for Intensify the power and capacity of the built system pattern in order to detect abnormal changes based on monitoring attacks, penetration attempts, and abnormal activities. And they give a detailed explanation according to the work that shows the ability to implement according to the NNs in establishing systems for monitoring abnormal changes (abnormalities) and the ability to differentiate between the normal pattern and the abnormal pattern.[35]

a different approach, Haripriya et al. (2018) develop a proposal through an updated taxonomy subject to ANN for intrusion detection using the backpropagation algorithm with the R tool. They employed feature selection techniques on the KYOTO dataset to improve performance measures such as F-measure, accuracy, and recall when compared to other models.[36]

Meanwhile, Wu et al. (2018) utilized a Convolutional Neural Network (CNN) for intrusion detection. This novel model addressed the imbalanced dataset issue and enhanced accuracy while reducing false alarm rates. Additionally, they introduced a model to convert raw traffic vectors into images, reducing computational costs. The proposed model was evaluated using the NSL-KDD dataset.[37]

On the other hand, Vinayakumar et al. (2017) opted for a Recurrent Neural Network (RNN) for their intrusion detection system. They surveyed large, many types of RNN and compared them to the patterns used by modular machine learning and advanced training. RNNs excel at learning temporal behaviors in large-scale sequence data. The authors applied their model to traffic data, particularly TCP/IP packets, from datasets such as DARPA, KDD-Cup-99, and UNSW-NB15, Powerful and intelligent detection of high frequency attack attempts such as DoS and Probe.[38]

Deep learning (DL) is a modern form of learning based on artificial neural networks. Work has been done with DL techniques within a wide range, noting unapproved traffic records, and working according to DL mechanisms in order to notice abnormal matters is far from our research, and it is planned to create deviant trend mechanisms based on DL to work in future matters.

Unsupervised Learning

Moving on to unsupervised learning techniques, these are clustering algorithms or undirected classification methods that do not require labelled data for training. Unsupervised methods aim to identify hidden patterns in data without using a pre-trained model. And it uses mechanisms to notice the abnormal situation according to the process of subjugation and control.

The K-means algorithm is a popular unsupervised clustering algorithm that divides observations into clusters based on similarity properties. Thakare et al. (2015) described the K-means algorithm and reviewed different approaches for outlier detection using this algorithm, emphasizing its applications in mining big data sets and stream data.[39]

Münz et al. (2007) proposed an anomaly detection method based on the K-means cluster algorithm for network data mining in the context of network security. This method divided unlabelled records into clusters of regular traffic and anomalies, using the K-means algorithm. The cluster centroids were used to detect anomalous traffic efficiently. The authors evaluated the method's effectiveness in detecting DoS attacks and port scans.[40]

While K-means is a fundamental clustering algorithm, its integration with other algorithms can enhance its effectiveness. Aung et al. (2018) presented a hybrid ML model that combined the K-means algorithm for identifying similar attack groups with a Random Forest algorithm to classify data as normal or attack. The proposed model achieved good results in detecting different types of intrusion attacks using the KDD-Cup-99 dataset.[41]

Hidden Markov Model – HMM

Hidden Markov Models (HMM) have gained widespread popularity in the field of data science and engineering as a state-based classification model. Initially used in speech recognition, HMM has proved to be successful in various analysis applications. Anomaly detection is one of the most critical applications of HMM, with numerous studies demonstrating its efficacy in this area. This paper focuses on recent research articles that employ HMM for security and intrusion detection purposes.

Chen et al. (2016) proposed an algorithm capable of handling large-scale data and event logs while recognizing the temporal relationship of unusual events. They also presented a state-based detection approach for identifying multi-stage advanced attacks. The primary challenge they faced was dealing with the enormous volume of data and devising effective methods for its analysis in the context of security. Results indicated that their proposed model performed successfully with a massive amount of event logs in the network.(42)

In the domain of Industrial Control Systems (ICS) security, Stefanidis et al. (2016) utilized HMM for intrusion detection. Specifically, they applied HMM to SCADA systems using interconnected TCP/IP protocol. To evaluate their system, they compared its detection accuracy with other existing systems that used the same datasets. The proposed system exhibited a higher detection rate for most attack vectors, and the researchers concluded that it was particularly suitable for real-time systems and high-speed environments.(43)

Addressing security concerns in 5G networks, Zegeye et al. (2019) developed a novel multi-layer approach based on HMM to protect networks against intruders and identify multi-phase attacks. They employed the CICIDS2017 dataset and applied techniques such as Singular Value Decomposition (SVD) and feature selection to reduce the data. K-means clustering labels were then used in monitoring the multi-layer HMM model. The proposed model demonstrated stable and well-trained performance, indicating that it did not require a large amount of training data.(44)

Meanwhile, securing mobile networks presents unexpected challenges, prompting researchers to develop models that can effectively overcome these challenges. According to Lian et al. (2018), the traditional HMM algorithm used for predicting network security lacks precision. To address this issue, they introduced a weighted HMM-based algorithm that specifically predicts mobile networking security. They employed multiscale entropy to overcome the slow training speed of data in the mobile networking domain while optimizing the HMM transition matrix. Additionally, they utilized the autocorrelation coefficient to establish the relationship between data characteristics and predict future network security. The algorithm's effectiveness was verified by implementing it on the DARPA2000 dataset, which contained various types of attacks, extensive data, redundancies, and false alarm rates. The experimental results demonstrated the accuracy and validity of the proposed model.(45)

In conclusion, HMM has found wide-ranging applications in the field of security and intrusion detection. Researchers continue to innovate and develop models that effectively utilize HMM to address the challenges posed by different domains such as network security and mobile networking. These studies highlight the effectiveness of HMM in detecting anomalies and protecting systems against various types of attacks.

Principal Component Analysis – PCA

Is a popular dimensionality reduction technique in the field of machine learning and data analysis. Its main objective is to simplify the complexity of high-dimensional datasets by transforming them into a lower-dimensional space without losing much information.

PCA works by finding the directions (principal components) in the original dataset that explain the maximum variance. These principal components are vectors that are orthogonal to each other. The first principal component captures the direction with the most variance, and subsequent components capture the remaining variance in decreasing order.

To perform PCA, the mean of each feature in the dataset is subtracted, and the resulting data is then transformed using linear algebra techniques. This process allows us to represent the data in a new coordinate system defined by the principal components.

PCA has several applications, including data visualization, noise reduction, and feature extraction. It helps to identify the most important variables in a dataset, enabling better understanding and interpretation of the data. Additionally, PCA can be used as a preprocessing step for other machine learning algorithms, as it reduces the dimensionality of the data and improves computational efficiency.

PCA is a powerful technique for reducing the dimensionality of high-dimensional datasets while retaining their essential characteristics. By extracting the most informative features, it provides insights and simplifies the analysis of complex data.

Ding et al. (2016) conducted experiments using PCA to detect anomalies in a network's traffic data. They successfully detected anomalies caused by node disconnection and DDoS attacks in a backbone network.(46)

Vasan and Surendiran (2016) evaluated the efficiency of PCA for anomaly detection and introduced the concept of Reduction Ratio. They found that the first 10 principal components were effective for classifying anomalies in different datasets. They concluded that using PCA in intrusion detection systems could improve accuracy and reduce complexity.(47)

Paffenroth et al. (2018) proposed a new anomaly detection system called Robust PCA, which used network packet data. The system achieved a low false positive rate and successfully detected various network attacks. It accurately detected previously unknown attacks, such as packet stream attacks.(48)

Hoang and Nguyen (2018) applied PCA to intrusion detection in IoT networks. They developed a new method for calculating distances and conducting anomaly detection in IoT networks using PCA. Their experiments on Kyoto Honeypot dataset showed quick detection and reduced computational complexity.(49)

Gaussian Mixture Model (GMM)

Gaussian Mixture Models (GMMs) have gained attention in recent research papers due to their diverse applications, such as network traffic authentication, traffic classification and authentication, outlier detection, and anomaly detection systems. These papers have highlighted the ability of GMMs to enhance various aspects of performance in these domains.

In the context of network traffic authentication, GMMs have shown promise in accurately identifying and authenticating network traffic. By modeling the probability distribution of network traffic data using a combination of Gaussian components, GMMs can effectively differentiate between legitimate and malicious traffic, thereby improving network security.

Traffic classification and authentication is another area where GMMs have demonstrated their potential. GMMs can be trained on labeled traffic data to classify and authenticate different types of network traffic accurately. This enables network administrators to identify and differentiate between various applications or protocols, facilitating efficient network management and resource allocation.

Moreover, GMMs have been employed for outlier detection in various domains. By learning the underlying distribution of a dataset, GMMs can identify data points that deviate significantly from the expected behavior, flagging them as potential outliers. This capability is particularly useful in detecting anomalies or abnormal patterns in network traffic, assisting in identifying potential security threats or system malfunctions.

Overall, these recent papers showcase the versatility and effectiveness of GMMs in several domains. By leveraging the strength of probabilistic modeling and the flexibility of mixture models, GMMs offer valuable insights into network traffic dynamics, contribute to accurate traffic classification, enhance authentication mechanisms, and enable the detection of outliers and anomalies. Future research in this area is likely to continue exploring and expanding the applications of GMMs in various performance-enhancing domains.

Lalitha and Josna (2016): This study utilized GMM for network traffic authentication. The model performed effectively, enhancing response time and packet delivery ratio, without impacting network performance.(50)

Alizadeh et al. (2015): The authors employed an unsupervised GMM approach for creating application models. The model accurately depicted actual traffic and was successful in identifying abnormal application traffic in multi-network environments.(51)

Reddy et al. (2017): This research utilized GMMs for outlier detection in univariate network traffic. The proposed methodology efficiently provided necessary information and showed promise in identifying outliers in different types of datasets and big data scenarios.(52)

Blanco et al. (2019): The authors utilized GMMs to model individual characteristics in a dataset as normal. Their approach outperformed other proposals and enhanced the performance and quality of anomaly detection systems.(53)

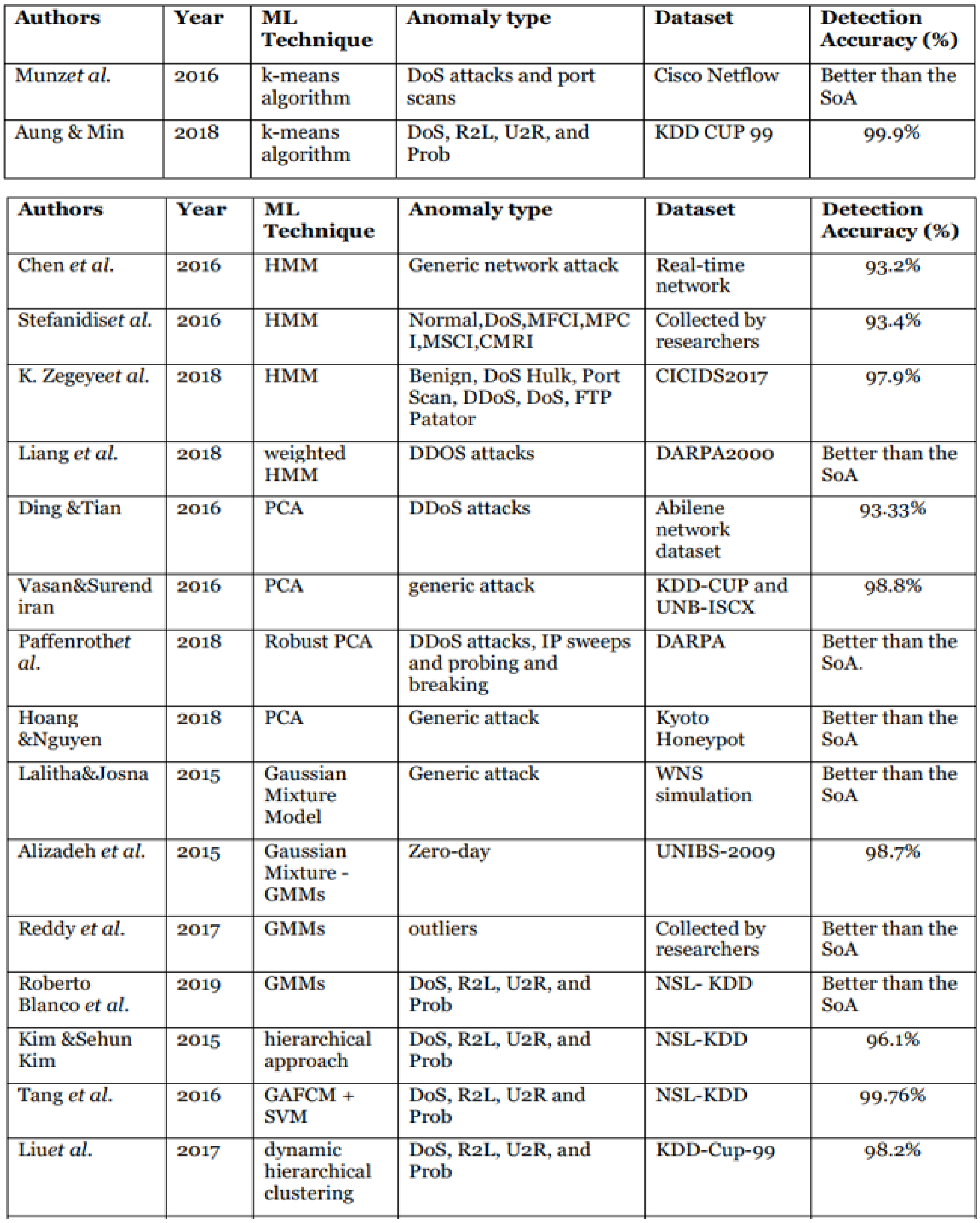
Hierarchical Clustering Algorithm

A method of grouping similar objects together to form clusters. It creates a hierarchy of clusters, with each cluster being different from another cluster and each object within a cluster being significantly similar to one another.

Kim et al. (2015) introduced a novel approach to Intrusion Detection Systems (IDS) by incorporating hierarchical clustering. By combining both misuse detection and anomaly detection models, their aim was to improve the detection rate of IDS while reducing the computational burden. the study by Kim & Kim demonstrates the potential of utilizing hierarchical clustering in IDS to optimize performance. By combining different detection models and leveraging a dataset for evaluation, they provide insights into enhancing the overall effectiveness and efficiency of intrusion detection systems.(54)

Tang et al. (2016) introduces an interesting approach to intrusion detection by combining fuzzy clustering and support vector machines in a hierarchical framework. The use of a hierarchical approach is beneficial as it allows for a more flexible and accurate detection approach, resulting in a high rate of correctly identifying intrusions while minimizing false alarms. Furthermore, the incorporation of support vector machines in the system helps in reducing the time required for training the model, which is an important practical consideration. Overall, this study appears to present promising results and contributes to the field of intrusion detection.(55)

Liu et al.'s (2017) novel and promising dynamic hierarchical clustering method. They effectively reduce feature dimensions through information gain-based feature selection. The use of generalized Euclidean distance to measure cross-domain data is clever. The incorporation of dynamic clustering accuracy as a guide is valuable. Their successful development of an anomaly detection model using this approach is impressive. Experimentation on the KDD-Cup-99 datasets further demonstrates the effectiveness of their approach, with high detection rates and low false alarms.(56)



Authors	Year	ML Technique	Anomaly type	Dataset	Detection Accuracy (%)
Munzet *al.*	2016	k-means algorithm	DoS attacks and port scans	Cisco Netflow	Better than the SoA
Aung & Min	2018	k-means algorithm	DoS, R2L, U2R, and Prob	KDD CUP 99	99.9%
Chen *et al.*	2016	HMM	Generic network attack	Real-time network	93.2%
Stefanidiset *al.*	2016	HMM	Normal,DoS,MFCI,MPCI,MSCI,CMRI	Collected by researchers	93.4%
K. Zegeyeet *al.*	2018	HMM	Benign, DoS Hulk, Port Scan, DDoS, DoS, FTP Patator	CICIDS2017	97.9%
Liang *et al.*	2018	weighted HMM	DDOS attacks	DARPA2000	Better than the SoA
Ding &Tian	2016	PCA	DDoS attacks	Abilene network dataset	93.33%
Vasan&Surendiran	2016	PCA	generic attack	KDD-CUP and UNB-ISCX	98.8%
Paffenrothet *al.*	2018	Robust PCA	DDoS attacks, IP sweeps and probing and breaking	DARPA	Better than the SoA.
Hoang &Nguyen	2018	PCA	Generic attack	Kyoto Honeypot	Better than the SoA
Lalitha&Josna	2015	Gaussian Mixture Model	Generic attack	WNS simulation	Better than the SoA
Alizadeh *et al.*	2015	Gaussian Mixture - GMMs	Zero-day	UNIBS-2009	98.7%
Reddy *et al.*	2017	GMMs	outliers	Collected by researchers	Better than the SoA
Roberto Blanco *et al.*	2019	GMMs	DoS, R2L, U2R, and Prob	NSL- KDD	Better than the SoA
Kim &Sehun Kim	2015	hierarchical approach	DoS, R2L, U2R, and Prob	NSL-KDD	96.1%
Tang *et al.*	2016	GAFCM + SVM	DoS, R2L, U2R and Prob	NSL-KDD	99.76%
Liu*et al.*	2017	dynamic hierarchical clustering	DoS, R2L, U2R, and Prob	KDD-Cup-99	98.2%

Figure 5. Unsupervisedanomaly detection approaches (SoA: State-of-the-art)

Comparison Between Supervised and Unsupervised Techniques

In Support Vector Machines (SVM), the combination of feature selection and parameter optimization has been shown to reduce both training and testing time, while also improving the effectiveness of the SVM classifier (Li et al., 2015). Similarly, in the Naïve Bayesian model, the integration of Principal Component Analysis (PCA) for feature extraction improves upon the traditional Naïve Bayesian algorithm by accounting for attribute weights.(57,58)

Previous studies presented in this survey indicate that supervised methods are commonly employed when working with non-real-time training data, due to their simplicity and efficiency. However, more flexible methods with a higher detection rate for known attacks are also utilized. Ensemble methods, which involve combining multiple classifiers, have shown to perform well even if individual classifiers are weak. Nonetheless, supervised methods have certain disadvantages, including resource consumption and time complexity when dealing with big data. Additionally, achieving real-time performance can be challenging.(59)

Unsupervised learning, on the other hand, eliminates the need for training data and is primarily utilized for feature detection. Unsupervised techniques aim to identify hidden patterns in data without the use of training data, enabling them to detect unknown attacks. For example, hierarchical clustering using the Fuzzy C-Means

https://doi.org/10.56294/dm202272

approach incorporates a membership function and fuzzy interval, allowing for the detection of unknown attacks (Yuan et al., 2011). Similarly, robust Principal Component Analysis (PCA) models have been successful in detecting anomalies/attacks that were not encountered or trained.[60,61]

The implementation of unsupervised learning techniques spans various areas and applications such as IoT, WSN, 5G mobile networks, and Industrial Control Systems (ICS). These applications often involve real-time data processing, and unsupervised techniques offer advantages such as fast response time and reduced computational complexity when dealing with large datasets. Unsupervised techniques can also achieve good accuracy results when combined with other classifiers in real-time networks. However, a significant limitation of anomaly detection is the detection rate, which is dependent on proximity measures and directly affects the false alarm rate. Time consumption is another challenge that future anomaly detection systems need to address.[62]

Semi-Supervised Learning

Semi-supervised learning is a machine learning approach that falls between the realms of supervised and unsupervised learning. It is employed when a dataset contains limited labeled examples and a large amount of unlabeled data. The goal of semi-supervised learning is to utilize the available labeled data along with the unlabeled data to improve the model's performance Aissa et al. (2016).[63]

The overall approach in semi-supervised learning involves leveraging the abundance of unlabeled data to obtain additional information about the underlying structure of the dataset. This additional knowledge can then be used to enhance the model's learning process and make better predictions.

There are several methods commonly used in semi-supervised learning. One approach is to initialize the model using the labeled data and then propagate the information from labeled to unlabeled data points. This propagation can be achieved through techniques like label propagation or graph-based methods.

Another approach is co-training, where multiple models are trained on different views of the data. Each model then labels the unlabeled data based on its own perspective, and the agreement between the models is used to provide labels for the unlabeled instances Ashfaq et al. (2017) [64]

The benefits of semi-supervised learning are numerous. It allows for the exploitation of large amounts of unlabeled data, which is often easier and cheaper to acquire compared to labeled data. This enables the model to learn from a more comprehensive representation of the dataset and potentially improve its accuracy and performance Borghesi et al. (2019) The experimental results showed that the autoencoder-based method significantly outperformed the supervised method, with a 12 % increase in accuracy.[65]

semi-supervised learning can be particularly useful in scenarios where obtaining labeled data is challenging or expensive, such as in medical diagnosis or natural language processing tasks. By incorporating the limited labeled data with the vast amount of unlabeled data, semi-supervised learning offers a powerful approach to address these challenges Yuan et al. (2016) achieved high accuracy rates of 93,71 %, 99,88 %, and 98,23 %, respectively.[66]

However, it is important to note that semi-supervised learning does not guarantee improved results in all scenarios. The effectiveness of this approach greatly depends on the quality and distribution of the available labeled and unlabeled data. Additionally, the choice of the specific semi-supervised learning algorithm or method employed can also impact the performance outcomes.

In conclusion, semi-supervised learning is a valuable technique that bridges the gap between supervised and unsupervised learning. It harnesses the potential of both labeled and unlabeled data to improve model performance and address challenges associated with limited labeled data. By effectively leveraging unlabeled data, semi-supervised learning can be a powerful tool in various domains and applications of machine learning.

CONCLUSION

The use of machine learning in anomaly detection for network traffic has proven to be an effective approach for early threat detection. By employing various algorithms and techniques, it is possible to train models to identify abnormal behavior and potential threats in real-time.

This technology offers several benefits, such as detecting previously unseen and sophisticated attacks, reducing false positives, and enabling prompt response and mitigation measures. Additionally, machine learning can continuously learn and adapt to evolving threats, making it a valuable tool in the ever-changing landscape of cybersecurity.

However, it is important to note that machine learning models are not a silver bullet and should not be solely relied upon for network security. They should be used as a complementary tool alongside other security measures to provide a layered defense approach. As the field of machine learning continues to advance, further research and development are needed to enhance the accuracy and efficiency of anomaly detection systems. Additionally, collaboration between cybersecurity professionals, data scientists, and machine learning experts is crucial to ensure the effective deployment and optimization of these models. Considering the ever-increasing sophistication of cyber threats, early threat detection is of paramount importance for organizations to safeguard their network infrastructure. The use of machine learning in anomaly detection for network traffic provides a promising solution to address this challenge and enhance overall cybersecurity defenses.

REFERENCES

1. Aburomman AA, Reaz MBI. A survey of intrusion detection systems based on ensemble and hybrid classifiers. Computers & Security. 2017;65:135-152.

2. Agrawal S, Agrawal J. Survey on anomaly detection using data mining techniques. Procedia Computer Science. 2015;60:708-713.

3. Ahmad S, Lavin A, Purdy S, Agha Z. Unsupervised real-time anomaly detection for streaming data. Neurocomputing. 2017;262:134-147.

4. Aissa NB, Guerroumi M. Semi-supervised statistical approach for network anomaly detection. Procedia Computer Science. 2016;83:1090-1095.

5. Akhi AB, Kanon EJ, Kabir A, Banu A. Network Intrusion Classification Employing Machine Learning: A Survey [Doctoral dissertation]. Department of Computer Science and Engineering, United International University, Bangladesh; 2019.

6. Alizadeh H, Khoshrou A, Zuquete A. Traffic classification and verification using unsupervised learning of Gaussian Mixture Models. In: 2015 IEEE international workshop on measurements & networking (M&N). IEEE; 2015. p. 1-6.

7. Amangele P, Reed MJ, Al-Naday M, Thomos N, Nowak M. Hierarchical Machine Learning for IoT Anomaly Detection in SDN. In: 2019 International Conference on Information Technologies (InfoTech). IEEE; 2019. p. 1-4.

8. Anderson JP. Computer security threat monitoring and surveillance. Technical Report, Fort Washington, PA, James P. Anderson Co; 1980.

9. Ashfaq RAR, Wang XZ, Huang JZ, Abbas H, He YL. Fuzziness based semi-supervised learning approach for intrusion detection system. Information Sciences. 2017;378:484-497.

10. Aung YY, Min MM. An analysis of K-means algorithm-based network intrusion detection system. Advances in Science, Technology and Engineering Systems Journal. 2018;3(1):496-501.

11. Bauer FC, Muir DR, Indiveri G. Real-Time Ultra-Low Power ECG Anomaly Detection Using an Event-Driven Neuromorphic Processor. IEEE Transactions on Biomedical Circuits and Systems. 2019;13:1575-1582.

12. Bhati BS, Rai CS, Balamurugan B, Al-Turjman F. An intrusion detection scheme based on the ensemble of discriminant classifiers. Computers & Electrical Engineering. 2020;86:106742.

13. Bhattacharyya DK, Kalita JK. Network anomaly detection: A machine learning perspective. CRC Press; 2013.

14. Blanco R, Malagón P, Briongos S, Moya JM. Anomaly Detection Using Gaussian Mixture Probability Model to Implement Intrusion Detection System. In: International Conference on Hybrid Artificial Intelligence Systems. Springer; 2019. p. 648-659.

15. Bock T. Displayr blog. https://www.displayr.com/what-is-hierarchical-clustering/

16. Borghesi A, Bartolini A, Lombardi M, Milano M, Benini L. A semi-supervised autoencoder-based approach for anomaly detection in high performance computing systems. Engineering Applications of Artificial Intelligence. 2019;85:634-644.

17. Chakir EM, Moughit M, Khamlichi YI. An effective intrusion detection model based on SVM with feature selection and parameters optimization. Journal of Applied Information Technology. 2018;96(12):3873-3885.

18. Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. ACM Computing Surveys (CSUR). 2009;41(3):1-58.

19. Chauhan P, Shukla M. A review on outlier detection techniques on data stream by using different approaches of K-Means algorithm. In: 2015 International Conference on Advances in Computer Engineering and Applications. IEEE; 2015. p. 580-585.

20. Chen CM, Guan DJ, Huang YZ, Ou YH. Anomaly network intrusion detection using hidden Markov model. International Journal of Innovative Computing, Information and Control. 2016;12:569-580.

21. Chew YJ, Ooi SY, Wong KS, Pang YH. Decision Tree with Sensitive Pruning in Network-based Intrusion Detection System. In: Computational Science and Technology. Springer; 2020. p. 1-10.

22. Rincon Soto IB, Sanchez Leon NS. How artificial intelligence will shape the future of metaverse. A qualitative perspective. Metaverse Basic and Applied Research. 2022. 27];1:12. https://doi.org/10.56294/mr202212.

23. Ding M, Tian H. PCA-based network traffic anomaly detection. Tsinghua Science and Technology. 2016;21(5):500-509.

24. Dua S, Du X. Data mining and machine learning in cybersecurity. CRC Press; 2016.

25. Duong NH, Hai HD. A semi-supervised model for network traffic anomaly detection. In: 2015 17th International Conference on Advanced Communication Technology (ICACT). IEEE; 2015. p. 70-75.

26. Fernandes G, Rodrigues JJ, Carvalho LF, Al-Muhtadi JF, Proença ML. A comprehensive survey on network anomaly detection. Telecommunication Systems. 2019;70(3):447-489.

27. Gu J, Wang L, Wang H, Wang S. A novel approach to intrusion detection using SVM ensemble with feature augmentation. Computers & Security. 2019;86:53-62.

28. Han X, Xu L, Ren M, Gu W. A Naive Bayesian network intrusion detection algorithm based on Principal Component Analysis. In: 2015 7th International Conference on Information Technology in Medicine and Education (ITME). IEEE; 2015. p. 325-328.

29. Haripriya LA, Jabbar M, Seetharamulu B. A Novel Intrusion Detection System Using Artificial Neural Networks and Feature Subset Selection. International Journal of Engineering and Technology. 2018;7(4):181.

30. Hu J, Ma D, Liu C, Shi Z, Yan H, Hu C. Network Security Situation Prediction Based on MR-SVM. IEEE Access. 2019;7:130937-130945.

31. Idhammad M, Afdel K, Belouch M. Semi-supervised machine learning approach for DDoS detection. Applied Intelligence. 2018;48(10):3193-3208.

32. Karim S, Rousanuzzaman PAY, Khan PH, Asif M. Implementation of K-Means Clustering for Intrusion Detection. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 2019;5:1232-1241.

33. Kevric J, Jukic S, Subasi A. An effective combining classifier approach using tree algorithms for network intrusion detection. Neural Computing and Applications. 2017;28(1):1051-1058.

34. Khraisat A, Gondal I, Vamplew P. An anomaly intrusion detection system using C5 decision tree classifier. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer; 2018. p. 149-155.

35. Kim E, Kim S. A novel hierarchical detection method for enhancing anomaly detection efficiency. In: 2015 International Conference on Computational Intelligence and Communication Networks (CICN). IEEE; 2015. p. 1018-1022.

36. Kotu V, Deshpande B. Data Science: Concepts and Practice. Morgan Kaufmann; 2018.

37. Kumar DA, Venugopalan SR. A novel algorithm for network anomaly detection using adaptive machine learning. In: Progress in Advanced Computing and Intelligent Engineering. Springer; 2018. p. 59-69.

38. Kusyk J, Uyar MU, Sahin CS. Survey on evolutionary computation methods for cybersecurity of mobile ad hoc networks. Evolutionary Intelligence. 2018;10:95-117.

39. Lakhina A, Crovella M, Diot C. Diagnosing network-wide traffic anomalies. ACM SIGCOMM computer communication review. 2004;34(4):219-230.

40. Lalitha KV, Josna VR. Traffic verification for network anomaly detection in sensor networks. Procedia Technology. 2016;24:1400-1405.

41. Larriva-Novo XA, Vega-Barbas M, Villagra VA, Sanz Rodrigo M. Evaluation of Cybersecurity Data Set Characteristics for Their Applicability to Neural Networks Algorithms Detecting Cybersecurity Anomalies. IEEE Access. 2020;8:9005-9014.

42. Albarracín Vanoy RJ. STEM Education as a Teaching Method for the Development of XXI Century Competencies. Metaverse Basic and Applied Research. 2022;1:21. https://doi.org/10.56294/mr202221.

43. Liu Y, Xu H, Yi H, Lin Z, Kang J, Xia W, Shi Q, Liao Y, Ying Y. Network anomaly detection based on dynamic hierarchical clustering of cross domain data. In: 2017 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C). IEEE; 2017. p. 200-204.

44. Manasa KN, Padma MC. A Study on Sentiment Analysis on Social Media Data. In: Emerging Research in Electronics, Computer Science and Technology. Springer; 2019. p. 661-667.

45. Mehmood A, Mukherjee M, Ahmed SH, Song H, Malik KM. NBC-MAIDS: Naïve Bayesian classification technique in multi-agent system-enriched IDS for securing IoT against DDoS attacks. The Journal of Supercomputing. 2018;74(10):5156-5170.

46. Meng X, Mo H, Zhao S, Li J. Application of anomaly detection for detecting anomalous records of terrorist attacks. In: 2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA). IEEE; 2017. p. 70-75.

47. Mohammadi S, Mirvaziri H, Ghazizadeh-Ahsaee M, Karimipour H. Cyber intrusion detection by combined feature selection algorithm. Journal of information security and applications. 2019;44:80-88.

48. Tovar Claros BS. Importance of design and user experience (UX) in web development. Metaverse Basic and Applied Research. 2022;1:20. https://doi.org/10.56294/mr202220.

49. Münz G, Li S, Carle G. Traffic anomaly detection using k-means clustering. In: GI/ITG Workshop MMBnet. 2007. p. 13-14.

50. Paffenroth R, Kay K, Servi L. Robust pca for anomaly detection in cyber networks. ArXiv preprint arXiv:1801.01571. 2018.

51. Peng H, Sun Z, Zhao X, Tan S, Sun Z. A detection method for anomaly flow in software defined network. IEEE Access. 2018;6:27809-27817.

52. Pham NT, Foo E, Suriadi S, Jeffrey H, Lahza HFM. Improving performance of intrusion detection system using ensemble methods and feature selection. In: Proceedings of the Australasian Computer Science Week Multiconference. 2018. p. 1-6.

53. Rai A. Optimizing a New Intrusion Detection System Using Ensemble Methods and Deep Neural Network. In: 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI). IEEE; 2020. p. 527-532.

54. Rai K, Devi MS, Guleria A. Decision tree-based algorithm for intrusion detection. International Journal of Advanced Networking and Applications. 2016;7(4):2828.

55. Reddy A, Ordway-West M, Lee M, Dugan M, Whitney J, Kahana R, Rao M. Using Gaussian mixture models to detect outliers in seasonal univariate network traffic. In: 2017 IEEE Security and Privacy Workshops (SPW). IEEE; 2017. p. 229-234.

56. Rettig L, Khayati M, Cudré-Mauroux P, Piórkowski M. Online anomaly detection over big data streams. In: Applied Data Science. Springer; 2019. p. 289-312.

57. Shukur HA, Kurnaz S. Credit Card Fraud Detection using Machine Learning Methodology. International Journal of Computer Science and Mobile Computing. 2019;8:257-260.

58. Stefanidis K, Voyiatzis AG. An HMM-based anomaly detection approach for SCADA systems. In: IFIP International Conference on Information Security Theory and Practice. Springer; 2016. p. 85-99.

59. Swarnkar M, Hubballi N. OCPAD: One class Naive Bayes classifier for payload-based anomaly detection. Expert Systems with Applications. 2016;64:330-339.

60. Tang C, Xiang Y, Wang Y, Qian J, Qiang B. Detection and classification of anomaly intrusion using hierarchy clustering and SVM. Security and Communication Networks. 2016;9(16):3401-3411.

61. Chandran R. Human-Computer Interaction in Robotics: A bibliometric evaluation using Web of Science. Metaverse Basic and Applied Research. 2022;1:22. https://doi.org/10.56294/mr202222

62. Thakare YS, Bagal SB. Performance evaluation of K-means clustering algorithm with various distance metrics. International Journal of Computer Applications. 2015;110(11):12-16.

63. Vasan KK, Surendiran B. Dimensionality reduction using principal component analysis for network intrusion detection. Perspectives in Science. 2016;8:510-512.

64. Veselý A, Brechlerova D. Neural networks in intrusion detection systems. Agricultural Economics (Zemědělská ekonomika). 2009;55(12):156-165.

65. Vinayakumar R, Soman KP, Poornachandran P. Evaluation of recurrent neural network and its variants for intrusion detection system (IDS). International Journal of Information System Modeling and Design (IJISMD). 2017;8(3):43-63.

66. Wang L, Li J, Bhatti UA, Liu Y. Anomaly Detection in Wireless Sensor Networks Based on KNN. In: International Conference on Artificial Intelligence and Security. Springer; 2019. p. 632-643.

67. Weerasinghe S, Erfani SM, Alpcan T, Leckie C. Support vector machines resilient against training data integrity attacks. Pattern Recognition. 2019;96:106985.

FUNDING

No financing.

CONFLICT OF INTEREST

None.

AUTHORSHIP CONTRIBUTION

Conceptualization: Mohammed Hussein Thwaini.
Research: Mohammed Hussein Thwaini.
Methodology: Mohammed Hussein Thwaini.
Writing - original draft: Mohammed Hussein Thwaini.
Writing - revision and editing: Mohammed Hussein Thwaini.