BIG DATA ANALYTIC USING MACHINE LEARNING ALGORITHMS FOR INTRUSION DETECTION SYSTEM: A SURVEY

Article · August 2021

DOI: 10.24247/ijmperdjun2020575

4 authors, including:

Abdulnaser A. Hagar
Dr. Babasaheb Ambedkar Marathwada University
8 PUBLICATIONS **23** CITATIONS

SEE PROFILE

Ali A Al-Bakhrani
Dalian University of Technology
13 PUBLICATIONS **59** CITATIONS

SEE PROFILE

Dr.Bharti W Gawali
Dr. Babasaheb Ambedkar Marathwada University
35 PUBLICATIONS **187** CITATIONS

SEE PROFILE

International Journal of Mechanical and Production Engineering Research and Development (IJMPERD)
ISSN (P): 2249–6890; ISSN (E): 2249–8001
Vol. 10, Issue 3, Jun 2020, 6063–6084

BIG DATA ANALYTIC USING MACHINE LEARNING ALGORITHMS FOR INTRUSION DETECTION SYSTEM: A SURVEY

ABDULNASER A. HAGAR*, DEEPALI G. CHAUDHARY, ALI A. AL-BAKHRANI & BHARTI W. GAWALI

Department of Computer Science & Information Technology, Dr. Babasaheb AmbedkarMarathwada University, Aurangabad, India

ABSTRACT

Through the increasing amounts of data day by day, big data analytics has appeared as an important trend for many organizations. These collected data can have important information that possibly will be key to solving extensive problems, such as cybersecurity, healthcare, marketing, intrusion detection, and fraud. An intrusion detection system (IDS) work as observers and evaluates the data in order to detect at all intrusion that may occur in the network or system. The huge data that is a high speed, high volume, and variety of data produced through the network has prepared for the process of the data analysis to detect any attacks through traditional techniques that cause big challenges. Big data analytic uses in IDS to increase accuracy and more efficient for detecting the attacks. we present a survey of eight machine learning algorithms use for IDS, intrusion detection techniques, advantages and disadvantages of intrusion detection techniques, all types of an intrusion detection system, advantages and disadvantages for each type of IDS, and five datasets that use on the intrusion detection system. Moreover, the authors attempt to present a strong picture of algorithms and datasets that use for IDS in all aspects through their extensive survey.

KEYWORDS: Big Data Analytics, Intrusion Detection Systems, Machine Learning Algorithms & Dataset for IDS

Received: May 13, 2020; **Accepted:** Jun 03, 2020; **Published:** Aug 03, 2020; **Paper Id.:** IJMPERDJUN2020575

Original Article

1. INTRODUCTION

Big data is the huge data that is increase speedily and focus area of research and a lot of techniques and frameworks proposed recently from many researchers. Big data become significant for all organizations private or public have collection great amounts, which may contai n valuable information around security issues such as fraud detection, national intelligence, and cybersecurity [1].Big Data leads to change some of the present jobs and creating new jobs. Companies are looking for persons that have skills of big data; many academies are presenting new certificates in order to deliver students by the big data skills.Companies and governments are able to collect data from many resources to get the information about where you go, what you do, what your preferences, and who friends of you. However, these techniques may improve the service and increase income to companies. Some legal restrictions on big data companies for example Google and Facebook because of what can do with the data they collect [2].

IDS is a software or hardware observé to analyzes and detect attacks that may occur in the network or the system. In the IDS there are three approaches to detect attacks; anomaly-based detection, Signature-based detection, and Hybrid-based detection. Signature-based detection is planned to detect attacks that are known via the attack signatures. Signature-based detection is so effective technique for detecting the known attack that is already loaded to the database of IDS. Consequently, The Signature-based detection is deemed to be accurate at identifying any

attacks try of known attacks [3]. While Anomaly-based detection matches predefined profiles against recent user activities to detect irregular actions that may be intrusions. Anomaly-Based detection is effective in touching new attacks without the need to update. However, Anomaly-Based detection has high false-positive rates. To overcome this difficult Hybrid-based detection is a blend of two approaches of intrusion detection to overcome the disadvantages in one method and get the advantages of the two methods that are used[4].

Currently, information security researchers are focusing on security use joining together intrusion detection capabilities that are identified, prevent the attack, log possible incidents, and also send reports to administrators. Intrusion Detection and Prevention System (IDPS) guarantees the availability, protection, confidentiality, and integrity of the information system. [5, 6].

The purpose of this paper is to highlight Big Data Analytics for Intrusion Detection and survey eight machine-learning algorithms that are used for intrusion detection. Moreover, presented tools of big data processing, modes for detect intrusion, intrusion detection techniques (detection methodology), types of intrusion detection systems, and datasets for the intrusion detection system.

2. BIG DATA ANALYTIC FOR INTRUSION DETECTION

Big data analytic uses to extract the knowledge from the data and helps us in attack detection. This section includes eight machine learning algorithms that are used for Intrusion Detection. Moreover, show the tools of big data processing, modes for detect intrusion, and intrusion detection techniques (detection methodologies).

2.1 Machine Learning Algorithms used for Intrusion Detection

Between a lot group of Machine Learning Algorithms, we have surveyed eight major Algorithms that are used for intrusion detection as following: Knowledge Nearest Neighbors [7], Support Vector Machine [8], Bayesian Network [9], Principle Component Analysis [10], Decision Trees [11], Random Forest [12], Fuzzy Logic [10], Genetic Algorithm (GA)[13].

2.1.1K-Nearest Neighbors (K-NN)

K-Nearest Neighbors (K-NN) depends on distance measures. K-NN methods hold the whole sampling set and includes the information found in the set and as long as the wanted grouping for the respective item. The sampling set must be processed the distance between each item aimed at the classifying where the K closet passages in the sampling set are considered as the point at a far distance. However, the shortcoming of the K-NN is the similarity measure, this leads to misclassification of points on account of its inefficiency accurate calculating distance between them, although classifying small subset of the features [14].

Saleh et al. in [15] offers a Hybrid IDS (HIDS) with three main contributions. The first contribution is the Naïve Base Features Selection (NBFS) technique for dimensionality reduction with two submodules: Mutual Effect Identification (MEI) and Feature Effect Identification (FEI). FEI identifies the significance of a single feature while MEI identifies the mutual importance of a pair of features using the NB classifier in a trial and error method. The second contribution is the Optimized Support Vector Machine (OSVM) for outliers rejection, which is primarily learned using high descriptive examples of each class, and use to remove outliers from the training dataset and the last contribution is the prioritized KNN (PKNN) for classification. PKNN is the enhancement of KNN, which considers as the average distance from KNN to the input point to be classified.

Syarif et al. in [16] have proposed a model which uses RF-based binary PSO for features selection wherein, at each generation, it performed attribute selection using binary PSO, and at the PSO loop classification achieved using RF. KNN is used for the classification of traffic packets.

2.1.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) an algorithm uses in classification and regression. SVM starts to construct a decision boundary that has an extreme edge among the typical data set and the source. SVM is a classifier for finding a separate hyperplane in the features space between two classes is such a path, to the point that the distance between the hyperplane and the open data points of each class is commonly increased. This method is the most effective for classifying the samples from the NSLKDD dataset, such as classifier is crucial in the intrusion detection system [17, 18].

Jaswal et al. in [18] introduced a hybrid approach that uses the technique which is K-means, SVM, and association rule mining algorithm for implementing IDS. Initially, data redundancy is controlled by applying the K-means clustering algorithm. Later SVM applied on top of these clusters and finally association rule mining applied to classify the KDD'99 data instance into normal or anomaly.

Thaseen et al. in [19] have proposed an intrusion detection model that works on rank-based chi-square features selection techniques and multiclass SVM classifier. Special parameter tuning technic employed for setting the parameters of SVM using a validation dataset. This optimal SVM is used for the classification of traffic packets.

Chen et al. in [20] proposed a New Ensemble Clustering (NEC) approach for intrusion detection using DB scan, One SVM, Agglomerative Clustering and Expectation-Maximization methods, wherein each method take a specific subspace of the original dataset and comes with classified labels, finally all the resulted from labels are ensembled and evaluated using a voting model.

Saxena et al. in [21] propose an SVM-PSO method for intrusion detection. SVM-PSO uses standard PSO for selecting SVM parameters and binary PSO for selecting the best features subset and SVM used for the classification of labels.

2.1.3 Bayesian Network (BN)

Bayesian Network (BN) is generally used in the grouping issues. BN tries to group into a qualitative model and quantities model. The qualitative has spoken via an organized non-cyclic graph that edges shape importance relations between the factors they interface and whose node means to random factors that are in the problem domain. The quantitative piece consists of some arrangement of probability distributions for each node. BN can be used to detect anomalies in vessel track [9, 22].

2.1.4 Principle Component Analysis (PCA)

Principle Component Analysis (PCA) is used for traffic characterization of a network segment to create a digital signature. PCA is a statistical procedure that is used to lessening dimensional multivariate issues via examining the differences for each variable amongst all response measurements. Moreover, the response of information able to symbolized through a decrease fixed set of dimensions devoid of damage information [10]. PCA can analyze important information from logs of data to find the significant activities of time intervals amongst the data set then diminish them this leads to the new set that is able to characterize regular conduct in the network segments [23].

Peng et al. in [24] have offered the use of the PCA algorithm and Mini batch K-means algorithm to get the clustering method for IDS. First, they have applied PCA to reduce the data dimensionality, and on top, of these to cluster the data they use mini-batch k-means and to initialize the cluster centers use k-means++ too. Each iteration is updating clusters by a new random sample from the data.

Vasan et al. in [25] experimented on the effectiveness of PCA for intrusion detection. They have identified the optimal number of principal components as ten for intrusion detection. They found that the classification accuracy enhanced by PCA if the data is noise-free.

Varghese et al. in [26] experimented the effectiveness with to reduction two features techniques PCA and CFS on different machine learning algorithms (J48, RF, LibSVM, Bagging with REP Tree, PART and MLP) using NSL-KDD dataset. Their performance results in a conclusion that out of all classifiers RF gives good classification accuracy when applied on PCA.

Chabathula et al. in [27] have made an experiment on the effect of PCA with dimensionality reduction on many algorithms of machine learning. The algorithms like Random Forest Tree, SVM, voting features Interval, Naïve Bayes probabilistic classifier. Tree algorithms give the highest classification accuracy.

Chowdhury et al. in [28] proposed a method for detecting malware that uses data pre-processing followed by feature extraction using n-gram and PE followed by PCA based feature reduction methods for enhancing detection accuracy. ANN with feed-forward is applied as a classifier.

2.1.5 Decision Tree (DT)

Decision Tree algorithm in structure is the same as the tree has leaves which leads to divisions and grouping and in this way declare the conjunctions of highlights that produce those classifications. There are two types under the DT algorithms such as C4.5 and ID3. C4.5 and ID3 algorithms formulate DT from the prearrangement of training data through the utilize of the idea of data entropy [17].

Goeschel et al. in [29] proposed a model for reducing false positive by combining SVM, and Naïve Bayes and Decision Trees methods. Initially, SVM is used to classify KDDCUP 99 dataset to separate into two classes (attack and normal) and then all instances of attack are classified by DT into the known attack and unknown attack. Then an unknown attack submitted to Naïve Bayes and likeliness to the other attacks is determined. If the high similarity is found it indicates a true positive otherwise this new alert should be submitted to farther investigation.

Paulauskas et al. in [30] analyses the influence of pre-processing of data on attack detection through using many Machine Learning Algorithms like Rule-Based classifiers, Decision Tree, and Naïve Bayes on the NSL-KDD dataset.

2.1.6 Random Forest (RF)

Random Forest (RF) is a collaborative algorithm used to improve accuracy. RF consists of two stages classification and feature selection. RF can produce multiple decision trees from a random subset of data. There is a major advantage of RF. RF yields low classification errors if we compare it with traditional classification algorithms. RF is used to detect many types of attacks such as a probe, R2L, DOS, and U2R [12]. Moreover, RF has high accuracy in prediction so RF uses in Robot detection problem and can handle diverse bots. Furthermore, RF consumes the time of computational if it works on a huge data set and complex procedure [31].

Mustapha et al. in [32] used MLlib and Apache Spark in order to make a test for the performance of intrusion detection via using many machine learning algorithms, the Algorithms are SVM, DT, Naïve Bayes, and RF, applied on the UNSW-NB15 dataset. They found RF yields the best accuracy of 97.49%, sensitivity 93.53%, and specificity 97.75%. The Naïve Bayes algorithm gives the worst accuracy of 74.19%.

Gupta et al. in [33] presented a framework of intrusion detection using Spark-based. In the framework applied by two features selection algorithms, one is Chi-squared feature selection and the second is correlation feature selection. They use five ML Algorithms (Logistic Regression, SVM, Naïve Bayes, GB tree, and RF). On the way to evaluate algorithms performance applied on DARPA and NSL-KDD datasets. Moreover, the results display that the top of accuracy is RF classifier but in prediction time the worst, although Naïve Bayes in the accuracy is the worst while is fast in prediction and training time.

Farnaaz et.al in [12] built an IDS by applying RF classifier for detecting attack groups like DOS, R2L, Probe, U2R. They have initially pre-processing the data followed by feature subset selection by applying Symmetrical Uncertainty (SU) measure on his preprocessed data and applied RF classifier only on the selected feature subsets.

2.1.7 Fuzzy Logic (FL)

Fuzzy Logic (FL) is in an environment that is uncertain. It can make rational decisions, incomplete, and inaccurate. FL is used to detect anomaly in-network uses time interval. Furthermore, to calculate the value of the threshold they applied the exponentially weighted moving average techniques that represent more current higher weight [10].

Hajimirzaei et al. in [34] offered IDS based on a Artificial Bee Colony algorithms(ABC) MultiLayer Perceptron(MLP) for IDS based on and fuzzy clustering. Homogeneous subsets of the training data are processed through fuzzy clustering. Moreover, ABC is used to optimize the MLP parameters, and also ABC is used for final classification of attack and normal.

2.1.8 Genetic Algorithm (G A)

Genetic Algorithm (GA) is considered as one of the heuristic search algorithms used in artificial intelligence and computing. GA is a method used for finding optimized solutions.

Kannan et.al. in [13] focus on feature selection for selecting test features which will help at the construction of a good model for IDS. They have used the Correlation-based Feature Selection (CFS) technique by using a mathematical intersection principle-based GA as a heuristic search algorithm. Features selection afterward CFS uses the GA algorithm. The effectiveness of the features selection technique performed after pre-processing is tested by using J48 classifiers and Naïve Bayes with helping of NSL-KDD dataset and achieves a good accuracy of 96.06% with Naïve Bayes classifier.

Ferriyan et al. in [35] emphasis on applying GA for selecting optimal features. For this they prepared three training datasets: ON, RM, and OA based on the relevance of features w.r.t attacks. Optimal features are selected from the three datasets using GA with a one-point cross over. RF is applied to these selected features of the datasets for getting the classification of labels.

Ariafar et al. in [36] proposed a framework for detecting attacks on the network that uses K-means and DT methods. GA is used for optimizing the parameters (value of K and the number of runs) of K-means and the confidence parameter of DT. This optimized K-means can be used for grouping data into clusters. The new data with updated cluster

labels are submitted for classification using an optimized DT classifier for attack detection. Their study aimed to improve the performance of the NEC approach offered by Chen in [20].

2.2 Tools of Big Data Processing

There are many tools used in processing Big Data. In many domains, it helps in intrusion detection. The tools of processing Big Data such as a spark, Hadoop, Storm, Kafka, Flume, and Amazon Kinesis have techniques that use ML algorithms to solve the problems. In Big Data processing some of those tools are used in real-time for intrusion detection.

2.3 Modes for Detect Intrusion

The modes types that are usually used to detect the intrusion are of three types (supervised, Semi-supervised, and unsupervised) that aim to predict by using algorithms for the training and testing phase. [37].

2.4 Intrusion Detection Techniques (Detection Methodologies)

Many detection methodologies are used for intrusion detection such as (misuse, rule-based, anomaly-based, and stateful protocol analysis)

- Misuse: compare patterns of attack which are represented via signature. Misuse is also called signature-based and can be identified and analyzed specific patterns of behavior or events. [38].
- Ruled-based: need to make the decision based on the rule sets that are already defined via the experts of the domain but the network traffic is increasing this leads to time-consumption and difficulty in coding and finding the rule sets [39].
- Anomaly-based: it is behavior-based, takes the input from the logs that are created from the operating system(OS). Anomaly-based looks if any behaviors variations to catch as masquerading. Moreover, anomaly-based monitoring generates profiles and uses it to analyze and detect anomaly behavior [38].
- Stateful protocol analysis: it works to detect the changes in protocol state. Different from anomaly-based the stateful protocol analysis accepts determined widespread profiles that are generated by vendors or industry leader [40].

Table 1: Advantage and Disadvantages of Intrusion Detection Techniques

Detection Techniques	Advantages	Disadvantages
Misuse detection	Very low in Rate false. It is the most effective protection against attacks and malware that have already been detected, identified, and categorized. Can detect known attacks effectively. Low computational cost. High accuracy of detecting known attacks. It has the ability to detect a known attack that is preloaded into a database of the intrusion detection system.	Only detect attacks that are already configured. It cannot detect novel attacks. Required update the signatures continuously. Attackers are able to make some adjustments to prevent matching signatures of a known attack.
Rule-based	Known attacks can be simply detected.	Difficult and time consuming for coding and finding rule sets. Time wasting and difficult for finding and coding rule sets. Unable to detect unknown attacks.

Anomaly-based detection	Can detect actually a wide range of novel attacks. Can detect the attacks from the network. It has the ability to decrease the false rate of unknown attacks. Low false positives. To collect behavior it uses a statistical test. It depends on the abuse of the privileges of the OS. Can be cheap to deploy and monitor.	May miss known attacks. In dynamic environments, it's less effective. The accuracy of detection depends on the amounts of collected features or behavior. It is possible to miss the novel attacks if they don't sticks out alongside the observed dimension. To configure the profile needs more time this leads to time-consuming. The purity of training data (i.e. absence of attacks).
Stateful analysis	Identifies unexpected sequences of commands. Distinguishes unexpected sequences of commands. Adds stateful characteristics to regular protocol analysis.	Be not able to detect attacks that do not disturb the features of commonly accepted protocol behavior. The resource-intensive for analysis and protocol state tracking.

3. TYPES OF INTRUSION DETECTION SYSTEMS

IDS types presented by collecting recent researches relevant to the field. Our approach in this survey is to conduct an overall review of all types related to all applications and environments of IDS. Researches of IDS have made good progress. Several studies have described the types of IDS.

Bruno et al. in [41] offered an overview of IDS for IoT and classified the IDS based on the method of detection, the threat of security, IDS location, and strategy of validation. In addition, discussed the different potentials for each attribute, and described intrusion detection techniques for IoT.

Aumreesh et al. [41] offered a study of IDS. The authors argued the intrusion detection and the types of IDS. The research gives emphasis to the range of IDS types such as network, host, and hybrid IDS. The authors also described every single type of IDS.

Zouhair et al. [42] introduced a general idea of various intrusions detection in the cloud and some techniques of detection used on IDS. Moreover, showed an overview of the cloud computing types based on IDS and divided Cloud-based into four types.

S. Soniya and S. Maria [43] introduced an overview of IDS techniques and classification. The research classified IDS into two main types: HIDS and NIDS. The study also classifies techniques that are used in detecting attacks for securing the network from new attacks. IDS types classified to many types depend on the deployed platform to detect attacks and may depend on the input data which is collected from many resources such as audit log, system call, the application process, user or system activities, and network traffic to detect and analysis attack. Also, IDS categorizes based on attack type that is able to detected by each type.

Liao et al. [44] introduced a wide-ranging review of the IDS and classified into four classes: Host-based IDPS (HIDPS), Network Behavior Analysis (NBA), Wireless (WIDS), Mixed IDS(MIDS). Also compared the four types based on the following criteria: Detection scope, Component, and Network architecture of each class. Table 2 shows the details of the four classes and a comparative between them.

Table 2: Comparative of Four IDS Types [44]

IDS	Components	Network Architecture	Detection Scope
HIDPS	Agent.Management server.Database server.	Managed networks or standard networks.	The host.
NIDS	Sensor:(inline/passive).Manage ment erver.Database server.	Managed networks.	Network or Host.
WIDS	Sensor:(passive).Management server.Database server.	Managed networks or standard networks.	WLAN.WLAN client.
NBA	Sensor: (most passive).Management server.	Managed networks or standard networks.	Network or host.

We can summarize the types of IDSs as the following :

3.1 Host based IDPS(HIDPS)

In many application levels and operating systems, it uses Host-based IDPS technology to monitor and detects the events on a host. HIDPS can use to analyze system settings such as (local security, software calls, policy, and audits logs) in the host for suspect activities. We can divide the functionality into HIDPS into four categories(File System Monitoring, Log File Analysis, Connection analysis, Kernel-Based HIDPS) [6].

The first developed type of intrusion detection was HIDS. HIDS analyzes and monitors the internal computing system or system-level activities of the single host such as application activity, system configuration, system logs or audit log, wireless network traffic (only for that host) or network interface, file access, running user or application processes and modification. The capabilities of HIDS contain event correlation, integrity checking, log analysis, rootkit detection, policy enforcement, hard-disk, memory, processor and battery utilization, and alerting. HIDS has a tendency to be less false positive and more accurate the network-based IDS for the reason that it analyses the log files, and as result, it can decide whether an attack successfully occurred or not [45, 46].

The host-based detection requires programs to be installed for the system to generate reports indicating if any malicious activity occurred. The problem with host-based systems is that they tend to be resource-intensive because they use the same computer resources installed on it and don’t have an operating system independent like other types of IDS [47]. There are many current systems that introduce host-IDS system types, for instance, OSSEC [48] and Tripwire [49]. HIDS can analyze the audit log files to identify and detect any intrusion system processes. Though, to analyze big amount of data to distinguish between malicious processes and normal processes needs lot of resources and long time of computation. A lot of researches introduced methods to solve this problem for instances.

Marteau [50] introduced new similarity measure in symbolic sequential data to detect an unknown attack. In the offered method the author focused on sequences of system calls through using the Sequence Covering algorithm for Intrusion Detection (SC4ID). SC4ID algorithm based on optimal-covering of a sequence by way of series of subsequences extracted from a predefined set of sequences.

Deshpande et al. [51] overview a model to analyze only selective system call traces to detect any malicious activities within the system and at that time alert the cloud user to found the malicious process.

Gautam et.al. in [52] proposed the Host-based Intrusion System Model (HISM) to detect intrusion using log files which are created through a single computer. They have used neural network models. They called the models as Multilayer Perceptron Neural Network (MPNN) and generalized Regression Neural Network (GRNN) and achieved high accuracies

with reduced FPR.

Subba et al. [53] overview framework to increase the efficiency of computation in HIDS. The introduced framework transformed the system call toward n-gram vector and reduced the input feature vectors size by the dimensionality reduction process. The feature vectors lastly analyze via many machine learning classifiers that named (C4.5 Decision Tree, SVM, MLP, and Naïve Bayes) to identify intrusive processes.

3.2 Network based IDS (NIDS)

Network-based IDS (NIDS) technology used on the application, transport, and network layer to analyze packets. NIDS is one of the most efficient technologies that are able to monitor and analyze real-time packets. NIDS unable to analyze traffic on mobile, mobile networks, and encrypted traffic, traffic [54].

NIDS is used to analyze and monitor the traffic of networks to specific network segments for suspicious activity detection. NIDS used in the packet-level analysis for all systems in the network segment by check IP, transport-network, and application protocol level activities and headers of the packet to detect various IP-based DOS attacks such as TCP SYN attack and fragment packet attack [55]. NIDS focuses more on the abuse of vulnerabilities whereas HIDS center around abuse of privilege [47]. NIDS quicker and costs less in response than HIDS for the reason that there is no need to maintain sensor programming at the host level, and it monitors traffic on close real-time or on real-time [56].

Consequently, NIDS can detect attacks as they occur. Though, NIDS does not indicate if such attacks are successful or not since it doesn't analyze encrypted network traffic to detect an attack, and it has restricted visibility inside the host machine [55].

Thus, until now, many pieces of research increased to develop effective methods for NIDS to detect attacks. Some products for network intrusion detection exist, such as NetSTAT [57], and Snort [58] which is a tool aimed at real-time NIDS.

Until now there are a lot of researches that introduced methods for NIDS such as:

Parvat et al. [59] proposed NIDS using deep learning. The proposed method has been used multiple binary classifiers which deep learning model through conquer and divide strategy. To evaluate the system NSL-KDD dataset has been used.

Suad et al. [60] introduced an intrusion detection model on a big data environment using a machine learning algorithm for feature selection used Chi-selector and for classification used SVM to reduce dimensionality in network traffic. To evaluate the proposed model KDD dataset has been used.

Sklavounos et al. [61] introduced a new method of NIDS for DOS attack detection depend on the exponential weighted moving average (EWMA) chart and on tabular cumulative sum (CUSUM) chart on the UDP and ICMP source bytes of the experimental dataset NSL-KDD.

3.3 Hybrid based IDS or Mixed IDS (MIDS)

MIDS combined two or more types of IDS to complete an accurate detection and attain the advantages of IDS such as Double Guard that uses HIDS and NIDS. Though, MIDS takes along time in analyzing data.

Tesfahun et al in [62] propose hybrid intrusion detection approach that detects both known and unknown attacks using two layers. The first layer implements signature-based using blocks detected (known) attack instances and RF classifiers. The second layer implements anomaly-based IDS by applying an ensemble of one-class SVM classifiers using bootstrap aggregating technique on the normal instances filtered out from the first layer. The detected attacks from this second layer are again blocked and updated to the train set.

Chiba et al. in [63] proposed a hybrid and cooperative NIDS that can detect both known and unknown cloud-attacks. CH-NIDS applied snort on the network packets for detecting known attacks as a first phase of detecting unknown attacks, it applies optimized Back Propagation Neural Network (BPNN) on the undetected packets of phase. CHNIDS should be deployed at the frontend and backend of the cloud for effective detection.

3.4 Wireless IDS (WIDS)

Wireless IDS (WIDS) is a type of NIDS that can be analyzed and monitor protocols and packets on a wireless network. Although WIDS is able to analyze the traffic of the network, unable to detect anomalous activities in applications [15]. WIDS analyses and monitor wireless traffic to detect any attacks. There are many types of attacks in a wireless network such as Spoofed altered routing attack, Sinkhole attack, and Sybil attack. Wireless networks have many features such as limited in computational power of sensors, existences in the open environment, battery life, and memory capacity; therefore, IDS that works on wired networks unable to be work totally on wireless networks. WIDS has more attacks than wired networks since their infrastructures are dynamic by nature [44], [64]. There are a lot of researches that introduced methods for WIDS such as:

Kolias et al. [65] introduced a distributed NIDS for wireless networks. The system is dependent upon swarm intelligence principles and classification rule induction to analyzed data for intrusion detection. For the test proposed method, they used the Aegean wireless intrusion dataset version2.

Gupta et al. [66] proposed a game theory on 5 G wireless cell access points to detect bandwidth spoofing attacks and analyzed the effect of it. The authors furthermore proposed WIDS used a model of hidden Markov to detect attacks and consider and focus the security issues of the 5 G wireless.

Patel et al. [67] summarized some of the advantages and disadvantages of the wired and wireless network. The author summarized the purpose of applying the IDS for WSN is to detect whether the node from the network is physical damage or malicious. WIDS significant to provide confidentially, integrity, and availability; therefore, it evaluates the signal jamming and eavesdropping.

3.5 Distributed IDS (DIDS)

Distributed IDS (DIDS) has numerous IDS and can communicate with a central server or with each other and allows network monitoring [6]. DIDS designed to work in a not homogenous environment this means that DIDS provides the capability to collect information from different sources to detect attacks against a network system for example Distributed Denial of Service (DDoS) attack or doorknob attack. DIDS has three components in the framework, which are communication component, IDS agent, and central analysis server. DIDS has many advantages compared to the centralized IDS showing in table 3 [68, 69]. There are many types of researchers that introduced DIDS such as:

Zeeshan and Peter [70]introduced and evaluated IDS methods on IoT, that are suitable for small devices. Moreover, the devices can manage the reputation data of neighbors in good management methods. The proposed methods made it probable to single out any behavior of malicious in an energy-friendly and processing way.

Arshad et al. [71] introduced a collaborative ID framework for Machine to Machine (M2M) based IoT, that leverages collaboration between IoT nodes for effective ID without consuming high communication, energy resources, and processing. The proposed framework planned the collective use of information from NIDS and HIDS. Also planned to address challenges for example the resource limitations of the nodes, flexibility, and the collaborative nature of the M2M networks.

Collaborative IDS is DIDS that has the capability to connect alarms coming from diverse sensors. Moreover, increase the potential to make the IDS autonomous, parallel, self-adjusting capabilities, organized and efficient. Isolated IDSs cannot achieve connections between malicious action happening at different places at the same time [72], [73].

Steven R. Snappl et al. [74] which introduced a prototype DIDS that generalized the target environment because monitor numerous hosts interconnected via network also the network itself.

Table 3: Distributed and Centralized IDS Advantages and Disadvantages

IDS	Advantages	Disadvantages
Distributed IDS	Scalability and flexibility. Detects DoS attacks for high-speed networks. Reduce computational costs. Analysis, monitoring, and processing of attack data are speedier and easier. Make conceivable an early intrusion detection that can result in blocking incoming traffic from specific IP addresses into the whole network.	The data stream between the agent and the host may possibly produce high network traffic overheads. The data whose path is long from its source to the IDS potentially modified or intercepted which possibly will result in misinterpretations. Can generate diverse outputs from different IDs.
Centralized IDS	The maintenance and administration cost lower compared to the case of distributed system. All the activities of IDS are controlled directly by the central console.	Not able to detect malicious events occurring at different places at the same time. IDS may unreliable or unusable because hackers possible can incapacitate the programs running on the system.

3.6 Network Behavior Analysis (NBA)

Network Behavior Analysis (NBA) is on types from NIDS. Moreover, can detect unusual activities that might come from any malware intrusion [6]. NBA can checks and monitors network traffic toward know attacks that produce uncommon traffic flows, such as malware, DDoS attack, and policy violations. The NBA system is able to investigate the network traffic to identify attacks using unexpected traffic flows [4, 44, and 75]. There are many NBA systems types of research such as :

Koch et al. [76] introduced a new NBA by used insider activities and similarity measurements such as data exfiltration in encrypted environments. In the proposed architecture used intrasession and intersession correlation, on the way to determine the similarity between connections.

Kakuru [77] introduced a tool for an internal network that suggestions a method to detect any uncommon behavior via an authentic user. In the proposed tool used Wireshark to record log traffic over a network. Initial, during a period of time the Wireshark recorded the user's activity and stored the record in a database. Afterward, the new activity is compared

to past activity and alerts any new behavior toward the administrator.

The advantage of the NBA is that it emphasizes the overall behavior of the devices on the network; therefore, it is allowed to respond to specific threats or unknowns for which no signature is available and zero-day attack.

3.7 Database IDS

Database IDS monitors and checks the attacks toward the database. There are types of database attacks for example direct DB attack and SQL injection attack [78]. Some researches addressed the SQL injection attack, for instance.

Liu A et al. [79] proposed SQL Proxy-based Blocker(SQLProb). In the SQLProb introduced method harnessed the Genetic Algorithms to extract user's entries and dynamically detect for adverse SQL, and used a proxy that integrated with environment presenting protection to frontend web servers and back-end databases.

3.8 Hypervisor-based IDS (Virtual Machine)

Virtual Machine or Hypervisor-based IDS Virtual IDS (V-IDS). VMI was introduced by Garfinkel et al. [80] by way of a hypervisor-level IDS which introduced isolation for the IDS, whereas still offering visibility into the state of the host. VM based IDS is progressive based on three VM abilities: Inspection, Isolation, and Interposition [81]. The hypervisor is a platform that runs VMs hypervisor-based IDS which works at the hypervisor layer. This help users to analyze and monitor the connections among VMs or amongst VM and hypervisor. It can preserve and also apply diverse security strategies to each VM based on their requirements. The most significant benefit of hypervisor-based IDS is the availability of information.

Table 4: Advantages and Disadvantages of IDS Types [54, 78, and 82]

IDS	Advantages	Disadvantages
HIDS	It is able to detect intrusion on network layer traffic, operating system, and host applications. In the system, the level does not permit intrusion to occur. It can analyze communications activity and encrypted data. It is able to detect any misuse of profiles. On the encrypted communication able to analyze and monitor suspicious activities. Able to inform us if an attack is successful or no. No need is required to added hardware because it works on the same host system.	The accuracy of detection is limited because HIDS does not use a predefined database. HIDS halt if the OS breakdown by the attack. It impacts system performance because it uses many host resources. It should be installed on every host. HIDS tends to be resource-intensive. It is monitored only by the host on which it is installed. HIDS are incapable to detect network scans or DoS attack.
NIDS	Has broad intrusion detection capabilities. The performance for hosts does not effect because of the Operating environment independent.	Does not indicate whether the attack was successful or no. Cannot analyze encrypted traffic. NIDS has limited visibility inside the host machine.
MIDS	More flexible. More efficient. It gathers the advantages of the strong point from combined types.	More overhead load because of combined methodologies. The utilization of processor in the hybrid agent is much great.
WIDS	It is effective for monitoring and analyzing intrusion on a wireless network. More accurate. WIDS can identify various problematic issues like misconfigurations and policy violations at the WLAN protocol level. It can manage wireless protocol activity.	Weak to DoS attacks. Sensors have limited energy and computational resources. The WIDS is susceptible to evasion techniques if attack channels are not presently monitored. It is unable to analyze and monitor packets on the network layer, transport layer, and application layer.

DIDS	More scalable than standalone IDSs. Processing, analysis, and monitoring of attack data is lower cost, easier, and speed.	Produces a high false alarm rate. Can have diverse outputs from different IDS.
NBA	Detect Dos attacks effectively. It is effective on a new attack that has no signature in the IDS database or detect zero-day exploits. Able to detect and monitor threats that come from malware, policy violation, reconnaissance scanning, reconstruct malware infections, and DoS attacks. In the network layer and transport layer the NBA efficient to monitor any packets.	It analyzes the packets in batches, this leads to delaying the rate to detect the intrusion. Some attacks may not be detected until they have already damaged systems especially attacks that occur quickly. Delay in detecting attacks.
V-IDS	Apply and preserve different security strategies for each VM. Offers a more robust view of the system.	Some virtualization systems make it easy to share information between the systems; this accessibility can turn out to be an attack vector especially if it is not carefully controlled. Virtualization adds additional layers that lead to increase in security management and controls overhead.

4. DATASETS FOR INTRUSION DETECTION SYSTEM

The main task of machine learning is to abstract the best value of information from data, so the performance methodology is understanding data. In IDS, the accepted data must be simple to reflect and acquire behaviors of networks or hosts. General sources of data types for IDS are flow, packets, logs, and sessions. It is difficult to build dataset and also take a long time. When the dataset built, many researchers can reuse it repeatedly.

4.1 DARPA1998

The Lincoln laboratory of MIT built the DARPA1998 dataset. DARPA1998 uses in IDS studies. It took from researchers nine weeks to collected internet traffic seven weeks for the training dataset then two weeks for the test dataset. DARPA1998 dataset consists of raw packets and labels. It contains five types of attacks Probe, Remote to Local (R2L), Denial of Service (DoS), User to Root (U2R), and Normal[83, and 84].

4.2 KDD99

The KDD99 dataset is one of the broadest used datasets in IDS datasets for the last period. It has 41-dimensional features take out from DARPA1998 and the labels (types of attacks) on KDD99 same as Labels (types of attacks) on DARPA1998. KDD99 contains four types of features (content, basic, time based statistical, and host-based statistical). Moreover, the KDD99 has a lot of redundant records and duplicated records so, it requires researchers some preprocessing before using the dataset such as filter the dataset. Finally, yet importantly, KDD99 is considered old to use in the current network environment. [84, and 85].

4.3 NSL-KDD

NSL-KDD proposed to overcome the disadvantages of the KDD99 dataset. In the NSL-KDD dataset, the process select record has done carefully. To avoid the bias problems of the classification in the NSLKDD records of different classes have well adjusted. Moreover, the NSL-KDD solves the redundant and duplicates records; thus, it only contains moderate records. Though, the NSL-KDD has not included in any new data; thus, it is still lacking minority class samples because the samples may be out of date. NSL-KDD dataset contains four types of attacks Probe, Remote to Local (R2L), Denial of

Service (DoS), and User to Root (U2R). In the following a brief description for the four attacks is mentioned:

- Denial of Service (DoS): is a type of attack that attempts to make the target system shut down traffic from and traffic to. In this case the system can't solve the problem and to protect the data the system does shut down which leads to the normal packets can't visit a network.
- Probe: is a type of attack that attempts to reach some information via the networks. The main aim of the Prob is to try to steal the target information from the networks.
- User to Root (U2R): is a type of attack. In this attack at the beginning, it works with an account of normal users to attempt to get access to the networks or the system.
- Remote to Local (R2L): is a type of attack attempts to get local access to the networks or the system[84].

4.4 UNSW-NB15

The University of South Wales built the UNSW-NB15 dataset to use by researchers on IDS and has nine types of attacks. The UNSW-NB15 dataset is suitable for new IDS based on machine learning. We will show these all nine types of attacks:

- Denial of Service: it is an intrusion. It disrupts the resources to make it unavailable to the user on the internet. It keeps the device extremely busy that even the authorized user will have no access.
- Exploits: the attacker has previous knowledge of the system or network and exploits the vulnerability.
- Reconnaissance: it is an attack that gets information about the system to get control of the system.
- Worm: it replicates itself with authentic user's information and gets into the system; spread itself and get access to the system.
- Fuzzers: this attack tries to acquire about security loopholes. It inputs enormous random data to crash the system.
- Analysis: this intrusion gets into a system through a web application through an open port.
- Backdoor: this type of attack bypasses the security mechanism of a system and gets access to the system or data.
- Shellcode: the attacker sends a small piece of code. It runs in a shell and becomes a control of the system.
- Generic: this is an attack on all kinds of block cipher [84, and 86].

4.5 CICIDS2017

The University of New Brunswick's (UNB) Canadian Institute for Cybersecurity (CIC) built the CICIDS2017 dataset. The CICIDS2017 contains 80 attributes and 3.1 million record and has seven attack types. The seven attacks are (Brute Force attacks, DoS attacks, hearable attacks, Botnet attacks, Infiltration attacks, and web attacks). The CICIDS2017 is a new dataset that uses by researchers for IDS [87].

5. CONCLUSIONS

Big data analytics can calculate the correlation between attributes from heterogeneous resources in IDS datasets to improve the security of the networks and hosts. We had introduced machine learning algorithms and big data analytics that use in IDS. We had survey recent works in big data analytic for IDS and the most eight algorithms that use for IDS. The IDS is

one of the most necessary considerations of cyber-security that can discover intrusion before and/or after an attack occurs. IDS plays an important role as a defense mechanism of networks and systems. IDS has improved dramatically over time, especially in the last few years due to the new advanced technologies. This paper also provides all types of IDS, the advantages and disadvantages of each type. We give a summary of IDS algorithms, techniques, methods, and datasets. All researchers can use this survey to create new algorithms, techniques, and tools to improve the performance of IDS and to increase the accuracy of detecting any attacks on hosts or networks.

6. REFERENCES

1. *J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," EURASIP J. Adv. Signal Process., vol. 2016, no. 1, 2016.*
2. *E. K. Clemons, J. Wilson, and F. Jin, "Investigations into consumers preferences concerning privacy: An initial step towards the development of modern and consistent privacy protections around the globe," Proc. Annu. Hawaii Int. Conf. Syst. Sci., pp. 4083–4092, 2014.*
3. *A. Sahasrabuddhe, S. Naikade, and A. Ramaswamy, "Survey on Intrusion Detection System using Data Mining Techniques," Int. Res. J. Eng. Technol., vol. 4, no. 5, pp. 1780–1784, 2017.*
4. *Chandra, Dimple, and Partibha Yadav. "Prediction Of Software Maintenance Effort On The Basis Of Univariate Approach With Support Vector Machine." International Journal Of Computer Science And Engineering (Ijcse) 3.3 (2014): 83-90.*
5. *J. Kizza, F. Migga Kizza, J. Kizza, and F. Migga Kizza, "Guide to Intrusion Detection and Prevention Systems," Secur. Inf. Infrastruct., pp. 239–258, 2011.*
6. *Jonathan Chee, "Host Intrusion Prevention Systems and Beyond," 2008.*
7. *K. Letou, D. Devi, and Y. Jayanta Singh, "Host-based Intrusion Detection and Prevention System (HIDPS)," Int. J. Comput. Appl., vol. 69, no. 26, pp. 28–33, 2013.*
8. *M. Wauters and M. Vanhoucke, "A Nearest Neighbour extension to project duration forecasting with Artificial Intelligence," Eur. J. Oper. Res., vol. 259, no. 3, pp. 1097–1111, 2017.*
9. *Deepa, S., and R. Umarani. "Steganalysis on images based on the classification of image feature sets using SVM classifier." International Journal of Computer Science and Engineering (IJCSE) 5.5 (2016): 15-24.*
10. *A. A. Aburomman and M. Bin Ibne Reaz, "A novel weighted support vector machines multiclass classifier based on differential evolution for intrusion detection systems," Inf. Sci. (Ny)., vol. 414, pp. 225–246, 2017.*
11. *S. Mascaro, A. Nicholson, and K. Korb, "Anomaly detection in vessel tracks using Bayesian networks," Int. J. Approx. Reason., vol. 55, no. 1 PART 1, pp. 84–98, 2014.*
12. *Chandolikar, N. S., and V. D. Nandavadekar. "Investigation of Feature Selection and Ensemble Methods for Performance Improvement of Intrusion Attack Classification." International Journal Of Computer Science And Engineering (IJCSE) Vol 2: 131-136.*
13. *A. H. Hamamoto, L. F. Carvalho, L. D. H. Sampaio, T. Abrão, and M. L. Proença, "Network Anomaly Detection System using Genetic Algorithm and Fuzzy Logic," Expert Syst. Appl., vol. 92, pp. 390–402, 2018.*
14. *A. P. Muniyandi, R. Rajeswari, and R. Rajaram, "Network anomaly detection by cascading k-Means clustering and C4.5 decision tree algorithm," Procedia Eng., vol. 30, no. 2011, pp. 174–182, 2012.*

15. N. Farnaaz and M. A. Jabbar, "Random Forest Modeling for Network Intrusion Detection System," Procedia Comput. Sci., vol. 89, pp. 213–217, 2016.

16. Danthala, S. W. E. T. H. A., et al. "Robotic Manipulator Control by using Machine Learning Algorithms: A Review." International Journal of Mechanical and Production Engineering Research and Development 8.5 (2018): 305-310.

17. A. Kannan, G. Q. Maguire, A. Sharma, and P. Schoo, "Genetic algorithm based feature selection algorithm for effective intrusion detection in cloud networks," Proc. - 12th IEEE Int. Conf. Data Min. Work. ICDMW 2012, pp. 416–423, 2012.

18. M. Y. Su, "Real-time anomaly detection systems for Denial-of-Service attacks by weighted k-nearest-neighbor classifiers," Expert Syst. Appl., vol. 38, no. 4, pp. 3492–3498, 2011.

19. A. I. Saleh, F. M. Talaat, and L. M. Labib, "A hybrid intrusion detection system (HIDS) based on prioritized k-nearest neighbors and optimized SVM classifiers," Artif. Intell. Rev., vol. 51, no. 3, pp. 403–443, 2019.

20. A. R. Syarif and W. Gata, "Intrusion detection system using hybrid binary PSO and K-nearest neighborhood algorithm," Proc. 11th Int. Conf. Inf. Commun. Technol. Syst. ICTS 2017, vol. 2018-Janua, pp. 181–186, 2018.

21. A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," IEEE Commun. Surv. Tutorials, vol. 18, no. 2, pp. 1153–1176, 2016.

22. K. Jaswal, P. Kumar, and S. Rawat, "Design and development of a prototype application for intrusion detection using data mining," 2015 4th Int. Conf. Reliab. Infocom Technol. Optim. Trends Futur. Dir. ICRITO 2015, 2015.

23. I. Sumaiya Thaseen and C. Aswani Kumar, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM," J. King Saud Univ. - Comput. Inf. Sci., vol. 29, no. 4, pp. 462–472, 2017.

24. W. Chen, F. Kong, F. Mei, G. Yuan, and B. Li, "A Novel Unsupervised Anomaly Detection Approach for Intrusion Detection System," Proc. - 3rd IEEE Int. Conf. Big Data Secur. Cloud, BigDataSecurity 2017, 3rd IEEE Int. Conf. High Perform. Smart Comput. HPSC 2017 2nd IEEE Int. Conf. Intell. Data Secur., pp. 69–73, 2017.

25. H. Saxena and V. Richariya, "Intrusion Detection in KDD99 Dataset using SVM-PSO and Feature Reduction with Information Gain," Int. J. Comput. Appl., vol. 98, no. 6, pp. 25–29, 2014.

26. J. Cózar, J. M. Puerta, and J. A. Gámez, "An application of dynamic Bayesian networks to condition monitoring and fault prediction in a sensored system: A case study," Int. J. Comput. Intell. Syst., vol. 10, no. 1, pp. 176–195, 2017.

27. G. Fernandes, L. F. Carvalho, J. J. P. C. Rodrigues, and M. L. Proença, "Network anomaly detection using IP flows with Principal Component Analysis and Ant Colony Optimization," J. Netw. Comput. Appl., vol. 64, pp. 1–11, 2016.

28. K. Peng, V. C. M. Leung, and Q. Huang, "Clustering Approach Based on Mini Batch Kmeans for Intrusion Detection System over Big Data," IEEE Access, vol. 6, no. c, pp. 11897–11906, 2018.

29. K. Keerthi Vasan and B. Surendiran, "Dimensionality reduction using Principal Component Analysis for network intrusion detection," Perspect. Sci., vol. 8, pp. 510–512, 2016.

30. J. E. Varghese and B. Muniyal, "An investigation of classification algorithms for intrusion detection system - A quantitative approach," 2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017, vol. 2017-Janua, pp. 2045–2051, 2017.

31. K. J. Chabathula, C. D. Jaidhar, and M. A. Ajay Kumara, "Comparative study of Principal Component Analysis based Intrusion Detection approach using machine learning algorithms," 2015 3rd Int. Conf. Signal Process. Commun. Networking, ICSCN 2015, pp. 1–6, 2015.

32. *M. Chowdhury, A. Rahman, and R. Islam, "Protecting data from malware threats using machine learning technique," Proc. 2017 12th IEEE Conf. Ind. Electron. Appl. ICIEA 2017, vol. 2018-Febru, pp. 1691–1694, 2018.*
33. *K. Goeschel, "Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and naive Bayes for off-line analysis," Conf. Proc. - IEEE SOUTHEASTCON, vol. 2016-July, 2016.*
34. *N. Paulauskas and J. Auskalnis, "Analysis of data pre-processing influence on intrusion detection using NSL-KDD dataset," 2017 Open Conf. Electr. Electron. Inf. Sci. eStream 2017 - Proc. Conf., pp. 1–5, 2017.*
35. *R. Genuer, J. M. Poggi, C. Tuleau-Malot, and N. Villa-Vialaneix, "Random Forests for Big Data," Big Data Res., vol. 9, pp. 28–46, 2017.*
36. *M. Belouch, S. El Hadaj, and M. Idlianmiad, "Performance evaluation of intrusion detection based on machine learning using apache spark," Procedia Comput. Sci., vol. 127, pp. 1–6, 2018.*
37. *G. P. Gupta and M. Kulariya, "A Framework for Fast and Efficient Cyber Security Network Intrusion Detection Using Apache Spark," Procedia Comput. Sci., vol. 93, no. September, pp. 824–831, 2016.*
38. *B. Hajimirzaei and N. J. Navimipour, "Intrusion detection for cloud computing using neural networks and artificial bee colony optimization algorithm," ICT Express, vol. 5, no. 1, pp. 56–59, 2019.*
39. *A. Ferriyan, A. H. Thamrin, K. Takeda, and J. Murai, "Feature selection using genetic algorithm to improve classification in network intrusion detection system," Proc. - Int. Electron. Symp. Knowl. Creat. Intell. Comput. IES-KCIC 2017, vol. 2017-Janua, pp. 46–49, 2017.*
40. *E. Ariafar and R. Kiani, "Intrusion detection system using an optimized framework based on datamining techniques," 2017 IEEE 4th Int. Conf. Knowledge-Based Eng. Innov. KBEI 2017, vol. 2018-Janua, pp. 0785–0791, 2018.*
41. *M. Kakavand, N. Mustapha, A. Mustapha, M. T. Abdullah, and H. Riahi, "A survey of anomaly detection using data mining methods for hypertext transfer protocol web services," J. Comput. Sci., vol. 11, no. 1, pp. 89–97, 2015.*
42. *I. Ghafir, M. Husak, and V. Prenosil, "A Survey on Intrusion Detection and Prevention Systems," Inf. Manag. Comput. Secur., vol. 18, no. 4, pp. 277–290, 2016.*
43. *A. G. Fragkiadakis, E. Z. Tragos, T. Tryfonas, and I. G. Askoxylakis, "Design and performance evaluation of a lightweight wireless early warning intrusion detection prototype," Eurasip J. Wirel. Commun. Netw., vol. 2012, pp. 1–18, 2012.*
44. *C. VARUN, B. ARINDAM, and K. VIPIN, "Anomaly detection A Survey," Comput. Mater. Contin., vol. 14, no. 1, pp. 1–22, 2009.*
45. *B. B. Zarpelão, R. S. Miani, C. T. Kawakani, and S. C. de Alvarenga, "A survey of intrusion detection in Internet of Things," J. Netw. Comput. Appl., vol. 84, pp. 25–37, 2017.*
46. *C. Zouhair, N. Abghour, K. Moussaid, A. El Omri, and M. Rida, "A review of intrusion detection systems in cloud computing," Secur. Priv. Smart Sens. Networks, pp. 253–283, 2018.*
47. *S. S. Soniya and S. M. C. Vigila, "Intrusion detection system: Classification and techniques," Proc. IEEE Int. Conf. Circuit, Power Comput. Technol. ICCPCT 2016, 2016.*
48. *H. J. Liao, C. H. Richard Lin, Y. C. Lin, and K. Y. Tung, "Intrusion detection system: A comprehensive review," J. Netw. Comput. Appl., vol. 36, no. 1, pp. 16–24, 2013.*
49. *L. Vokorokos and A. BaláŽ, "Host-based intrusion detection system," INES 2010 - 14th Int. Conf. Intell. Eng. Syst. Proc., pp. 43–47, 2010.*

50. P. Chauhan and N. Chandra, "A Review on Hybrid Intrusion Detection System using Artificial Immune System Approaches," Int. J. Comput. Appl., vol. 68, no. 20, pp. 22–27, 2013.
51. H. Kozushko, "Intrusion Detection : Host-Based and Network-Based Intrusion Detection Systems," Computer (Long. Beach. Calif)., 2003.
52. R. Bray, D. Cid, and A. Hay, OSSEC host-based intrusion detection guide. 2008.
53. G. H. Kim, E. H. Spafford, and W. Lafayette, "The Design and Implementation A File System Integrity of Tripwire : Checker."
54. P. F. Marteau, "Sequence Covering for Efficient Host-Based Intrusion Detection," IEEE Trans. Inf. Forensics Secur., vol. 14, no. 4, pp. 994–1006, 2019.
55. P. Deshpande, S. C. Sharma, S. K. Peddoju, and S. Junaid, "HIDS: A host based intrusion detection system for cloud computing environment," Int. J. Syst. Assur. Eng. Manag., vol. 9, no. 3, pp. 567–576, 2018.
56. S. K. Gautam and H. Om, "Computational neural network regression model for Host based Intrusion Detection System," Perspect. Sci., vol. 8, pp. 93–95, 2016.
57. B. Subba, S. Biswas, and S. Karmakar, "Host based intrusion detection system using frequency analysis of n-gram terms," IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON, vol. 2017-Decem, pp. 2006–2011, 2017.
58. N. A. Azeez, T. M. Bada, S. Misra, A. Adewumi, C. Van der Vyver, and R. Ahuja, "Intrusion Detection and Prevention Systems: An Updated Review," Adv. Intell. Syst. Comput., vol. 1042, no. January, pp. 685–696, 2020.
59. C. Modi, D. Patel, B. Borisaniya, H. Patel, A. Patel, and M. Rajarajan, "A survey of intrusion detection techniques in Cloud," J. Netw. Comput. Appl., vol. 36, no. 1, pp. 42–57, 2013.
60. W. Int, A. Computer, and A. C. Science, "Protocol-Based Classification for Intrusion Detection," pp. 6–11, 2008.
61. Vigna Giovanni and R. A. Kemmerer, "NetSTAT: A network-based intrusion detection approach," Proc. - Annu. Comput. Secur. Appl. Conf. ACSAC, pp. 25–34, 1998.
62. M. Roesch, "Snort – Lightweight Intrusion Detection for Networks," pp. 229–238, 2015.
63. A. Parvat, S. Dev, S. Kadam(B), and J. C. Sinhgad, Network Intrusion Detection System Using Ensemble of Binary Deep Learning Classifier, vol. 876. Springer Singapore, 2018.
64. S. M. Othman, F. M. Ba-Alwi, N. T. Alsohybe, and A. Y. Al-Hashida, "Intrusion detection model using machine learning algorithm on Big Data environment," J. Big Data, vol. 5, no. 1, 2018.
65. D. Sklavounos, "Utilization of Statistical Control Charts for DoS Network Intrusion Detection," Int. J. Cyber-Security Digit. Forensics, vol. 7, no. 2, pp. 166–174, 2018.
66. A. Tesfahun and D. L. Bhaskari, "Effective Hybrid Intrusion Detection System: A Layered Approach," Int. J. Comput. Netw. Inf. Secur., vol. 7, no. 3, pp. 35–41, 2015.
67. Z. Chiba, N. Abghour, K. Moussaid, A. El Omri, and M. Rida, "A Cooperative and Hybrid Network Intrusion Detection Framework in Cloud Computing Based on Snort and Optimized Back Propagation Neural Network," Procedia Comput. Sci., vol. 83, pp. 1200–1206, 2016.
68. I. Butun, S. D. Morgera, and R. Sankar, "A survey of intrusion detection systems in wireless sensor networks," IEEE Commun. Surv. Tutorials, vol. 16, no. 1, pp. 266–282, 2014.
69. C. Kolias, V. Kolias, and G. Kambourakis, "TermID: a distributed swarm intelligence-based approach for wireless intrusion detection," Int. J. Inf. Secur., vol. 16, no. 4, pp. 401–416, 2017.

70. *A. Gupta, R. K. Jha, and S. Jain, "Attack modeling and intrusion detection system for 5G wireless communication network," Int. J. Commun. Syst., vol. 30, no. 10, pp. 1–14, 2017.*
71. *A. Patel, M. Taghavi, K. Bakhtiyari, and J. Celestino Júnior, "An intrusion detection and prevention system in cloud computing: A systematic review," J. Netw. Comput. Appl., vol. 36, no. 1, pp. 25–41, 2013.*
72. *D. Kshirsagar, S. Sawant, R. Wadje, and P. Gayal, "Distributed Intrusion Detection System for TCP Flood Attack," pp. 979–989, 2017.*
73. *Z. C. Johany, "Survey on intrusion detection systems," pp. 1–5, 2015.*
74. *Z. A. Khan and P. Herrmann, "A trust based distributed intrusion detection mechanism for internet of things," Proc. - Int. Conf. Adv. Inf. Netw. Appl. AINA, pp. 1169–1176, 2017.*
75. *J. Arshad, M. M. Abdellatif, M. M. Khan, and M. A. Azad, "A novel framework for collaborative intrusion detection for M2M networks," 2018 9th Int. Conf. Inf. Commun. Syst. ICICS 2018, vol. 2018-Janua, pp. 12–17, 2018.*
76. *G. Folino and P. Sabatino, "Ensemble based collaborative and distributed intrusion detection systems: A survey," J. Netw. Comput. Appl., vol. 66, pp. 1–16, 2016.*
77. *E. Vasilomanolakis, S. Karuppayah, M. Muhlhauser, and M. Fischer, "Taxonomy and survey of collaborative intrusion detection," ACM Comput. Surv., vol. 47, no. 4, pp. 1–33, 2015.*
78. *S. R. Snapp et al., "DIDS (Distributed intrusion detection system) - Motivation, architecture, and an early prototype," Proc. 14th Natl. Comput. Secur. Conf., pp. 1–9, 1991.*
79. *FSabahi and AMovaghar, "Intrusion detection: A survey," Proc. - 3rd Int. Conf. Syst. Networks Commun. ICSNC 2008 - Incl. I-CENTRIC 2008 Int. Conf. Adv. Human-Oriented Pers. Mech. Technol. Serv., pp. 23–26, 2008.*
80. *R. Koch, M. Golling, and G. D. Rodosek, "Behavior-based intrusion detection in encrypted environments," IEEE Commun. Mag., vol. 52, no. 7, pp. 124–131, 2014.*
81. *S. Kakuru, "Behavior based network traffic analysis tool," 2011 IEEE 3rd Int. Conf. Commun. Softw. Networks, ICCSN 2011, pp. 649–652, 2011.*
82. *S. M. Othman, N. T. Alsohybe, F. M. Ba-alwi, and A. T. Zahary, "Survey on Intrusion Detection System Types," vol. 7, no. December, pp. 444–462, 2018.*
83. *A. Liu, Y. Yuan, D. Wijesekera, and A. Stavrou, "SQLProb: A proxy-based architecture towards preventing SQL injection attacks," Proc. ACM Symp. Appl. Comput., pp. 2054–2061, 2009.*
84. *T. Garfinkel and M. Rosenblum, "A Virtual Machine Introspection Based Architecture for Intrusion Detection," Ndss'03, vol. 1, pp. 253–285, 2003.*
85. *P. S. Martinez, "Virtual Machines and Security," pp. 1–7, 2013.*
86. *U. A. Sandhu, S. Haider, S. Naseer, and O. U. Ateeb, "A Survey of Intrusion Detection & Prevention Techniques," 2011 Int. Conf. Inf. Commun. Manag., vol. 16, pp. 66–71, 2011.*
87. *J. Mchugh, "Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPA Intrusion Detection System Evaluations as Performed by Lincoln Laboratory," ACM Trans. Inf. Syst. Secur., vol. 3, no. 4, pp. 262–294, 2000.*
88. *N. Moustafa and J. Slay, "A Comprehensive Data set for Network Intrusion Detection systems," 2015.*

89. *N. Moustafa, G. Creech, and J. Slay, "Big Data Analytics for Intrusion Detection System : Statistical Decision-Making Using Finite Dirichlet Mixture Models."*

90. *A. A. Shah, Y. D. Khan, and M. A. Ashraf, "Attacks Analysis of TCP And UDP Of UNCW-NB15 Dataset," VAWKUM Trans. Comput. Sci., vol. 15, no. 3, p. 143, 2018.*

91. *R. Panigrahi and S. Borah, "A detailed analysis of CICIDS2017 dataset for designing Intrusion Detection Systems," Int. J. Eng. Technol., vol. 7, no. 3.24 Special Issue 24, pp. 479–482, 2018.*

AUTHORS PROFILE

Mr. Abdulnaser A. Hagar Mr. Abdulnaser A. Hagaris a Lecturer in Al-Baydha University, Al-Baydha, Yemen. He is in present perusing Ph.D. under the guidance of Professor and Head Dr. Bharti Wamanrao Gawali in the Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS) India. Mr. Abdulnaser A. Hagar has completed Post Graduate degree (M.Sc.) Computer Information Systems, Middle East University for Graduate Studies, Faculty of Information Technology, Amman – Jordan with The first one in the Department and the Faculty of Information Technology.

Mr. Abdulnaser A. Hagar has completed Graduate degree (B.Sc.) Computer Information Systems from Applied Science University, Faculty of Information Technology, Amman-Jordan with Excellent with honor The first one in the Department and the Faculty of Information Technology. He has three research publications (Scopus) to his credits. His area of research interest includes Big Data Analytic, Cyber Security, Internet of Things and Attack Detection.

Ms. Deepali G. Chaudhary Ms. Deepali G. Chaudhary in present perusing Ph.D. under the guidance of Professor and Head Dr. Bharti Wamanrao Gawali in the Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS) India.

Ms. Deepali G. Chaudhary has completed Graduate degree (B. C. S.) in Computer Science from Dr. Babasaheb Ambedkar University, Aurangabad, Post Graduate degree (M.Sc.) in Computer Science from Dr. Babasaheb Ambedkar University, Aurangabad. M.Phil degree in Computer Science from Dr. Babasaheb Ambedkar University, Aurangabad.

She has 03 research publications to her credits.Her area of research interest includes Brain Computer Interface, 3D Modeling.

Mr. Ali A. Al-Bakhrani Mr. Ali A. Al-Bakhrani has completed Graduate Degree B.Sc. from the Faculty of Computer Sciences & Information Systems from THAMAR UNIVERSITY – Yemen.

M.Sc. Information Technology from Dr. BABASAHEB AMBEDKAR MARATHWADA UNVERSITY. He has three publications. His area of research interest includes Image Processing, and Data Science.

Professor Bharti Wamanrao Gawali Dr. Bharti Wamanrao Gawali in present is working as a Professor and Head in the Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS) India.

Dr. Bharti Gawali has completed Graduate degree (B. Sc.) in Computer Science from Dr. Babasaheb Ambedkar University, Aurangabad, with First Division, Post Graduate degree (M.Sc.) in Computer Science from Dr. Babasaheb Ambedkar University, Aurangabad, with First, Qualified SET in Computer Science and Application from University of Pune, and Doctoral Degree (Ph.D. (Data Compression)) in Computer Science from Dr. Babasaheb Ambedkar University, Aurangabad.

Professor Bharti Gawali has 21 years of experience in teaching, research and innovation.

Professor Bharti Gawali is a recipient of different awards such as Shikshak Pratibha award on 5th September 2009 by Department of Mass Communication and Journalism, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. Second Prize of Maharashtra Research festival (Avishkar) held at Health University, Nasik 2011. DST Fast Track Young Scientist award and Ideal Teacher award (Adarsh Shikshak Puraskar) on 5th September 2014 Dr. Babasaheb Ambedkar Marathwada University, Aurangabad.

She has 101 research publications to her credits. Out of which 48 research publications from SCI / Scopus / Web of Science Indexed Journals Peer reviewed Journals.

She has presented 21research papers in numerous International and National conferences. Till date she has delivered more than 41 invited talks at various.

Her h-index is 11 andi10-index is 15. Her work is being cited for632 times.

She has successfully supervised around 09 Ph.D.,17 M. Phil and 03M. Techscholars for their research. Currently 05 Ph.D., 2 M. Phil and 03M.Techstudents are working with her.

She has undertaken 3 Major Research projects (completed) and 1 Minor Researchprojectof Rs. 45, 29, 500 from various funding agencies via. Dr. B. A. M. University, UGC, DST and DST (SERB). She is currently working on 1 major project entitled as "Establishment of Science Technology and Innovation Hub for Empowerment of SC/ST Populations" of Rs. 2,61,86,209/- funded by DST-SEED STI Hub.

She has visited abroad for participating in BCI Workshop in Netherland.

Professor Bharti Gawali is actively associated with various National and International academic organizations. Life Member of IAEng, CSTA, IACSIT, IEEE, FIETE and ISCA.

Her area of research interest includes Data Compression, Speech Processing, Human Computer Interface, Emotion Recognition, Signature Recognition, Brain Computer Interface, Remote Sensing and GIS, Medical Image Processing etc.