Hindawi
Journal of Computer Networks and Communications
Volume 2018, Article ID 6383145, 10 pages
https://doi.org/10.1155/2018/6383145

Review Article

Role of Machine Learning and Data Mining in Internet Security: Standing State with Future Directions

Bilal Ahmad,[1] **Wang Jian,**[1] **and Zain Anwar Ali**[2]

[1]*Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, 29 Yudao St., Nanjing 210016, China*
[2]*Sir Syed University of Engineering & Technology, Karachi, Pakistan*

Correspondence should be addressed to Bilal Ahmad; ahmad@nuaa.edu.cn

Received 8 November 2017; Revised 26 February 2018; Accepted 28 March 2018; Published 2 July 2018

Academic Editor: Rui Zhang

As time progresses with vast development of information technology, a large number of industries are more dependent on network connections for sensitive business trading and security matters. Communications and networks are highly vulnerable to threats because of increase in hacking. Personnel, governments, and armed classified networks are more exposed to difficulties, so the need of the hour is to install safety measures for network to prevent illegal modification, damage, or leakage of serious information. Intrusion detection, an important entity towards network security, has the ability to observe network activity as well as detect intrusions/attacks. This study highlights the developing research about the application of machine learning and data mining in Internet security. We provide background, enthusiasm, discussion of challenges, and recommendations for the application of ML/DM in the field of intrusion detection.

1. Introduction

This study covers literature review of ML/DM techniques for cybersecurity applications. The complexity of different techniques, current achievement, and limitations in developing IDS is elaborated. Due to the effect of extraordinary approaches in information technology and large-scale usage of communication and Internet, people are motivated to transfer information using IT-based environment. This has advantages and benefits, like trimming the big graphical distance and exchange of information with ease. On the other side, such type of information transfer creates problems like intrusion and malicious activities which can disturb the communication. The security management becomes more difficult because of smart hacking techniques. The earlier safety trends of computer networks rely on stationary methods in which the OS need to be updated frequently for prevention of security dumps, and firewalls are also deployed at the serious network area to improve the security. The goals of firewalls are to adjust and control the flow of information inside and outside of a network rather than to detect whether or not the network is under attack. For balancing simple firewall, the intrusion detection system was used to collect and analyze network logs to predict possible security threats including intrusion-like outer organization attack and misuse detection-like attack from the organization. IDS has an important role in network security infrastructure to provide important security line, but unfortunately, most of these met a number of challenges like low detection rate and high false alarm rate; these problems are due to the complexity of the threats and have similarities to the normal behavior. IDS can be hardware or software or a combination of both, responsible to reveal the intrusion from network log data. An attempt of the intruder can be a sophisticated step to break the system security and initiated by collecting the information about the system like the used protocol and system available on network. Hackers start to probe every system to analyze different vulnerabilities; after the vulnerabilities are targeted, hackers try to get primary control by remote-to-local (R2L) attack. After user access, attacker tries to get opportunity by user-to-root (U2R) attack, if attacker gets master user access it can have privilege of stealing or modification, and if the targeted system is negotiated/compromised then the attackers have authority to go further at this step. To handle network security problems, many devices have been made during the last few decades, and IDS is one of them. IDS can be in the categories of host-based and network-based form where the host-based intrusion detection (HIDS) refers to intrusion detection that takes place on a single host system, and on the other hand, network-based intrusion detection system (NIDS) is installed on a certain place into the network to monitor network activities. The intrusion can have categories of three main types which include misuse-based/signature-based, anomaly-based, analytical-based/statistical-based, and hybrid-based techniques. Misuse-based techniques are used to detect already known attack from database signature. These types of techniques required regular updates of database containing signature and rules. Anomaly-based techniques observe any abnormal activity in the network and send alerts to the admin. These techniques worked on finding patterns of suspicious data. The suspicions attitudes can be anomalous or outliers. This technique is suitable for detecting the zero-day attack. The hybrid-based technique is the mixture of both techniques previously studied and shows that this technique is better than the other both, but this technique is complex. In this study, both anomaly and hybrid detection are discussed together. This study is useful for those who want to start research in the area of ML/DM for internet intrusion detection including some examples given as to how these techniques helped to improve cybersecurity. Bhuyan et al. provided an overview of the network anomaly detection method and network intrusion detection system by classifying network intrusion detection methods and NIDS into categories and also provided analysis of some techniques in the ways of their efficiency and ability [1]. Padhy et al. made a survey on data mining application, and according to their survey, many data mining techniques can be useful for anomaly detection [2]. Agrawal and Agrawal told that the categorized IDS depends on signature and anomaly-based detection techniques; for signature-based system, it needs to recognize the traffic patterns; patterns database need to be updated regularly. Anomaly-based detection matches or compares all activities for normal behavior [3]. Ahmed et al. presented a deep analysis of four important types of anomaly detection methods which are information theory, clustering, classification, and statistical analysis. They evaluated the techniques and explored the recently concerned publically available dataset [4]. Najeeb and Dhannoon surveyed the present feature selection methods, including wrapper, filter, and hybrid, and they analyzed and presented the performance of these methods for IDS with the KDDcup dataset and NSL-KDD dataset [5].

The focus of our research on ML and DM methods is highly compatible with intrusion detection with the cable-connected network.

2. Role of Machine Learning and Data Mining

The differences between the data mining, machine learning, and KDD is that KDD is the method to extract knowledge/useful information from the record. Data mining contains algorithms to recognize pattern from data. Availability of many research studies and results informed us that all KDD processes of DM defines KDD steps (preparation, selection, cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of DM) which used specific algorithms to extract patterns from record [6]. ML and DM are mostly confusing because of similar meanings, so they have a meaningful similarity. The founder of machine learning, Arthur Samuel, describes it as an area of research which gives the ability to learn without being explicitly programmed. ML interacts with learning the pattern recognition and computational learning theory in artificial intelligence. ML is older than DM. In the recent days, the term data mining is extra popular than its sibling machine learning which can be the reason for some scholars to actually highlight their study for data mining than machine learning, so in this study, machine learning and data mining are discussed together.

2.1. Training, Validation, and Testing Set. In machine learning, the following three steps are involved: training, validation, and testing. A training phase is part of data for exploration of possible analytical relationships, and the test part is the portion of data use to determine the strength and efficiency of relationship. In case of most appropriate classifier, the needed training set is used to train the algorithms, and the validation set has the ability to compare their efficiency and get the decision, which one to use, and then, finally the test set gets the performance like accuracy, specificity, and sensitivity. There are no specific criteria to determine how one could split the dataset; for example, it could be divided as 50 percent training, 25 percent validation, and 25 percent test. Data mining and machine learning are basically classified into supervised and unsupervised approaches. Figure 1 shows the supervised and unsupervised methods for ML.

In unsupervised learning work with nonlabeled data, the goal for unsupervised learning is to find knowledge from unlabeled data. If a part of data is labeled then the problem is known as semisupervised learning, and if all data are labeled then the problem is known as supervised learning.

2.2. PMML (Predictive Model Markup Language) for DM/ML. PMML gives an approach to analyze the application for description and discussion about the predictive model prepared by data mining and machine learning algorithms [7]. It is based on XML and capabilities for logistic regression, and the feedforward neural network advance version supports NB, KNN, and SVM classifiers. PMML was found in 2009, and until now, PMML has 4.0, 4.1 and 4.2, 4.2.1, and 4.3 versions.

2.3. CRISP Data Mining Model. This model provides a look of life sequence of data based on data mining as shown in Figure 2. This model is based on six basic phases [8]. *Business understanding*: In this earlier step, understanding the plan objectives and its needs from the business viewpoint is developed; after this, reshape the information into data mining problem solution to get an initial plan scheme. *Data understanding*: It starts with collecting data processes with

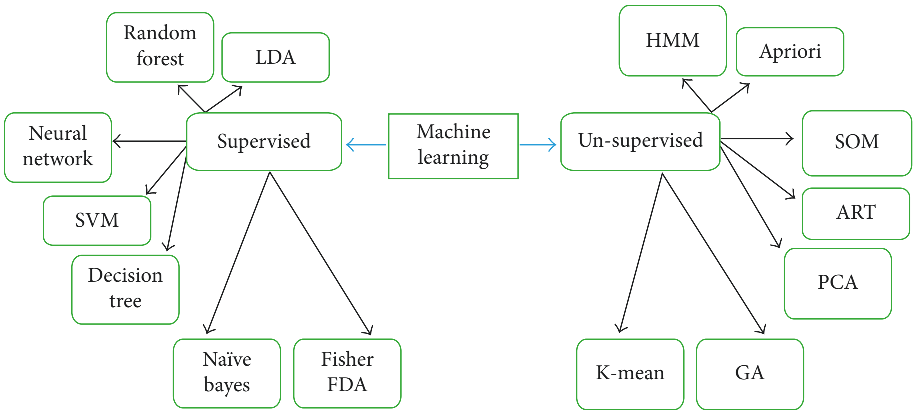


Figure 1: Supervised and unsupervised machine learning techniques.

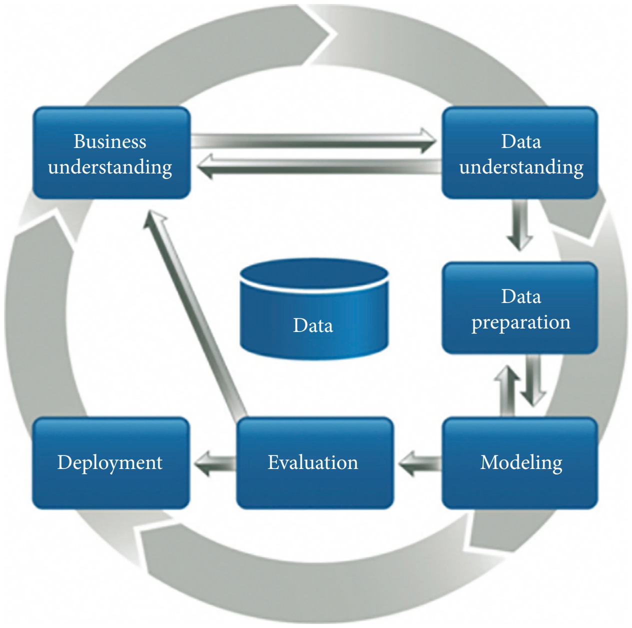


Figure 2: CRISP data mining model.

events to get awareness of data, highlights the problems, determines inside of data, or discovers existing subset for hypotheses of unseen problems. *Data preparation*: It is all about finalizing the dataset, and it performed many times. *Modeling*: Many modeling method selections are implemented in this step to evaluate the model to verify business aims. *Evaluation*: Before submission of the project to the final stage, it is required to evaluate to ensure that all important business issues should be completely addressed. *Development*: It can be as simple as production of a report or as complex as applying a data mining process. Mostly users can handle these steps of development mainly to recognize what steps will need to make in order to use of created models.

2.4. Evaluation Measures for DM/ML. Estimating the performance of data mining methods is the fundamental feature of machine learning. The most popular methods for performance measure are mentioned. Cross validation is the technique for assessing the simplification about presentation of the analytical model. Here, data are split once or many times for approximating the risk of each method. A portion of the data training set is used to train the algorithm, and the remaining data validation set is used to approximate the risk of the method. K-fold cross validation is responsible for dividing into k divisions, each time one of the k subsets is used as the test set, the remaining k-1 subset is placed together to make a training set, and then, the regular errors across all k trials are calculated.

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier"). In the field of ML, the confusion matrix also called the error matrix is a specific table layout which shows the performance of the algorithm especially for supervised learning. Terminologies and derivations from a confusion matrix are as follows: *Accuracy*: It measures the percentage of accurate estimation from the total number of instances calculated $\text{Acc} = (tp + tn)/(tp + fp + tn + fn)$. *Precision*: The helpful patterns which are properly estimated from the total number of estimations from the positive class $p = tp/(tp + fp)$. *Recall*: It calculates all correctly classified positive patterns $r = tp/(tp + tn)$. True negative ratio: It also known as specificity used in some area in fraction of negative patterns that rightly classify $sp = tn/(tn + fp)$. *Area under curve*: AUC gives the total ranking presentation of a classifier $\text{AUC} = (s_p - n_p(n_n + 1)/2)/n_p n_n$. *F-Measure*: The harmonic mean between recall and precision values $\text{FM} = (2 * p * r)/(p + r)$. There are two types of metric use for unsupervised learning techniques, which are internal and external: first, internal metrics are used for the cluster data, and second, external metrics perform a statistical testing on the structure of data.

For machine learning and data mining methods, datasets are important, and the following are some descriptions of few famous datasets related to Internet security or intrusion detection.

3. Intrusion Detection Dataset

3.1. Publically Available Dataset

3.1.1. DARPA 1998 for Intrusion Detection. The Lincoln Lab at the MIT University was supported by DARPA (Defense

Advanced Research Project Agency) and AFRL (Air Force Research Lab), to develop the dataset known DARPA used to detect intrusion and evaluation. This dataset contains network traffic just like real traffic between a little air force-base and the Internet masses for network IDS. The four main classes of attack added in the dataset are DOS, R2L, U2R, and probe [9]. DARPA 1998 dataset has two sets: first, seven weeks of training data with normal background traffic and labeled attacks and second, two weeks of test data with unlabeled data having new attacks.

DARPA 1998 has information for network connections, and each instance has 41 features which can be divided into four groups [10]: features 1 to 9 associate with "Basic Features," 10 to 22 with "Content Features," number 23 to 30 comes from "Time-Based," and 31 to 41 links to "Host-based Traffic features." Thomas et al. [11, 12] observed the usage of DARPA in 2008 after the ten years of dataset creation. They concluded that the dataset is not outdated and useful for IDS evaluation. It showed good prospect in attack on network traffic such like Dos, U2R, R2L, and probe. They concluded that this dataset can be presented as a baseline for any study or research. *KDDcup 1999 Dataset*: This is the most used dataset in examining IDS since 1999. It is based on DARPA 1998 for cybersecurity examination, and the training dataset has nearly about 4,900,000 connections with 41 features labeled as normal or attack with a specific category.

Corrected KDDcup99 dataset: This corrected KDDcup99 dataset can be used where all redundant connections were removed to adopt to these ways of probabilities of biasness of classifier can be reduced.

10 percent KDDcup dataset: The complete version of the dataset is not used often due to its large size for the training and test tasks. Mostly, 10 percent of the portion of this dataset is used. Training the classifier on the reduced dataset makes it possible computationally *NSL-KDD dataset*: This dataset is created from the KDDcup99 dataset in 2009; it contains 125,973 records for the training dataset, and the test dataset has 22,544 records.

DARPA 1998 dataset is categorized into 4 types of attack: denial of services (DoS), user-to-root (U2R), remote-to-local (R2L), and probe. In the Dos attack type, intruders attempt to get access to target the resources. In U2R type, the intruder attempts to get the root access. A R2L attack is an attempt to get access of the local network. In the probe attack type, the intruder collects knowledge about network resources.

Packet-Based Data: IETF (Internet Engineering Task Force) produced the IP (144) list which used large types of protocols like TCP, UDF, ICMP, and so on The application runs these and produces network packets. These packets sent and received via an Ethernet port etc., collected by the APIs (application programming interface) known as Libpcap and Winpcap are the examples for packet collection software libraries with protocol analyzer, sniffing, monitoring, and network IDS. At the physical network layer, the Ethernet border is composed of the Ethernet header (Mac address) with up to 1500 bytes of payload which contains the IP packet and composed of the IP/transport layer header and IP payload. The IP payload may have data or other summarized high grade protocols like NFS (Network File System), HTTP (Hypertext Transfer Protocol), POP (Post Office Protocol), and so on due to total packets collected by the pcap interface, and the structure of data varies in the sense of protocol which the packet carries.

Netflow Data: Basically, Cisco introduced Netflow. The switch has capacity to gather IP from network traffic. Netflow by Cisco version V explains network flow as a unidirectional order of packets which shares the right same VII packets features: source-IP, destination IP address, source port, destination port, IP-type of service, ingress interface, and IP-protocol. The basic Netflow structure has three modules: Netflow collector, Netflow Exporter, and Analysis Console. Presently, 10 versions of Netflow are available. Netflow data are the compressed and preprocessed form of real network packets.

4. Data Mining and Machine Learning Techniques for Intrusion Detection

This area has description of different data mining and machine learning techniques for intrusion detection. Every method with the description detail and reference of inspiring work is given.

4.1. Decision Tree.

The DT is similar to the flow down structure, and the construction of DT has a leaves representation of classification. Most famous types of DT are ID3 and C4.5, and Ross Quinlan work is based on these types of decision trees. ID3 is based on the concept learning system (CLS). ID3 is a supervised learning method and the tree is made on the basis of information gain that comes from training instances after using the same classification on the test data. ID3 mostly uses nominal attributes with zero missing values. It uses entropy and information gain to build a decision tree. C4.5 is the extension of the ID3 algorithm. The tree is simple; technically, it looks easy to have a tree that modifies the prospects of variables to be tested, and typically, a balanced tree can have good results.

Entropy is used to calculate the homogeneity of sample data. In the case of full homogeneity, the entropy will be zero, and if the sample is equally divided then entropy will be one. Information gain is based on reduction in entropy after data dived on an attribute. Building a decision tree is all about searching attributes that gives the highest information gain. The benefits of decision trees are pure knowledge expression and high accuracy, and the common disadvantage is that data with categorized variables have multiple levels, and information gain value may be biased due to the favor of feature with many levels. Decision tree information is based on the highest information gain, resulting in normal variable raking. Kruegel and Toth [13] used the decision tree model in snort. They presented an idea by using the misuse detection model in snort with decision tree. The experiment was performed by the DARPA 1999 dataset. The results showed that the clustering and decision tree can considerably reduce the processing time for the misuse-based detection system. EXPOSURE [14] is a mechanism that is based on wide scale, passive DNS analysis methods to detect the domains that have participation in suspicious activity. The systems have four basic parts: Data Collector, Feature Attribution Component, Malicious Benign-Domains Collector, and learning Module Classifier. Classifier made by J48 Decision Tree is available in the Weka application. By using 10-cross validation, accuracy of millions of domain was 98.5 percent and the FP rate was 0.9 percent. Further, the experiment was made to know whether the technique was successful to detect suspicious domain in which information was present not in the dataset. Initially, the experiment is performed from 50 randomly selected domains from 17686 domains and the classifier diagnoses three benign domains as suspicious with a 6-percent FP rate. After this experiment, another experiment was performed in which research validated the suspicious domain noticed by the classifier by using online-site-rating software like Norton Safe Web, Google Safe Browsing, and so on. The FP rate was noticed to be 7.9 percent for this study. In the third part, this study finds out 3117 new suspicious domains which were not used previously by the training set with any FP rate during the period. The results present accuracy and FAR of this research (EXPOSURE) and are acceptable and useful to identify malicious domain like botnet command scam hosts.

Aljawarneh et al. [15] proposed a research to enhance the decision tree (j48) method, and the experimental work included the NSL-KDD dataset. The result showed 99.88 percent detection accuracy with all features by using 10-fold cross validation. In the future work, they aimed at using the suggested algorithm in real network to further explore the j48 algorithm. Ahmad et al. [16, 17] proposed the two hybrid models to enhance intrusion detection by using the decision tree; in the first model, they used DT with the meta-algorithm to make high accuracy with a low false alarm rate. The experimental result involved the KDD dataset and showed that the proposed method is more effective than the other data mining algorithms. The second proposed method used DT with KNN, and the suggested method showed a good detection result for probe, U2R, R2L, and Dos attacks.

4.2. Artificial Neural Networks. Artificial neural network inspired by the human brain includes huge number of neurons with every neuron having an input and output with an activation function. Commonly, neural network is based on the layer approach; the first layer is the input layer, the last layer is the output layer, and the other layers are known as hidden layers as shown in Figure 3.

Dias et al. [18] proposed an idea by using ANN with the KDDcup intrusion detection dataset. The result of the experiment showed that the proposed idea can get 99.9 percent accuracy which is very much satisfactory than the previously developed methods.

Sodiya et al. [19] proposed a unique idea by using a combination of self-organization map and multilayer perceptron. The experiment showed that the proposed system improved accuracy up to 4 percent and detection up to 96 percent with 3 percent false alarm. For some people, the improvement is little small, but in the area of intrusion detection, one single improvement can have an ability to threaten the security of all networks. Vinchurkar and Reshamwala [20] made a review about neural network with machine leaning techniques to enhance the security of network by determining the network behavior. The study described the neural network approach as well as machine learning methods to face the challenges for IDS which also includes the study of SVM for classifier building problems, and in the future direction, there is a need to develop a system which has the ability to handle the recent challenges of IDS with a high detection ratio and performance.

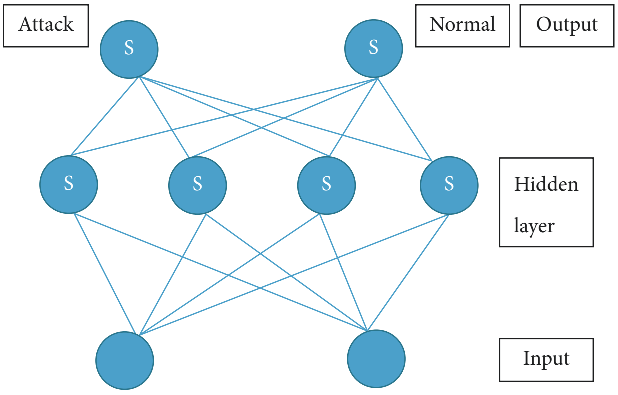


Figure 3: Representation of a perception with one hidden layer.

4.3. Association Rules/Fuzzy Association Rules. The rule-based method is used for exploring useful relation between variables in large data. It is based on which variable helps to discover association between apparently unconnected data in information source. Brahmi et al. [21] applied the association rule on the DARPA dataset which observed the relationship between TCP/IP and attack types. The study explained the multidimensional association rule where more than one precursor was present in the rules. For example, in the four-dimensional rule, IF (service AND src_port AND dst_port AND num_conn), THEN attack_type. Best presentation of working was found by using the six-dimensional rule having a detection rate of 99, 95, 75, and 87 percent for attacks (DoS, probe, U2R, and R2L). This experiment is not recommended for higher dimensions due to cost problems. Nalavade and Meshram [22] presented an idea which integrates association rules for the network intrusion detection system, and the suggested methods made attack rules for anomaly-attack detection. The suggested study by use of the KDD dataset showed that the modification in association rules is capable to detect intrusions in the network.

4.4. Bayesian Network (BN). Bayesian network is an analytical graphical model which shows the random variable, and their dependencies, for example, can show the statistical relationship among diseases and indications [23]. BN is made by use of high information or use of the expert algorithm which performs implication. In the case of anomaly-based detection, it is assumed to be sensitive as the system answers to an input if the input is unforeseen. Equally misuse-based detection is problematic, as the system is positive due to the signature that comes from the input are being scanned always against the attack patterns. Choosing the positive method as in misuse detection needs categorizing of the network traffic. Devarakonda et al. [24] proposed a study based on Bayesian network and the Hidden Markov Model with help of the KDDcup dataset as intrusion detection dataset. The model for IDS has been developed with different stages like learning the model with the training dataset and construction of Bayesian network, and this arrangement is used as the HMM stage diagram. The model trained and tested for normal and attacked data recodes one by one. The experimental results showed the high classification for normal and attacked data. Alocious et al. [25] used the Bayesian network classification technique for anomaly-based detection. They used this classification model to build a strong and accurate intrusion detection system. By using the KDD dataset, they used four different Bayesian network for classification, where every model connected and joined with each other to forecast network traffic data. The suggested model showed the capacity to detect novel attacks by ongoing learning with high accuracy and efficacy to detect attack data in the test dataset. The given study also has a proper approach to train and detect attacks from live network flow. Goeschel [26] promised to develop the system that has the ability to reduce the false positive ratio. They make combination of NB, SVM, and DT to improve the overall accuracy of IDS and efficiency. The suggested system was divided into three steps with each step having the role of the mentioned classifier. The proposed IDS showed 99.62% true positive results, and the author suggested his model over the other network as a future work.

4.5. Clustering. The clustering approach is used to group the specific set of items which depends on the upload of their characteristics, collecting them with regard to their resemblances. Basic benefit of the clustering technique in intrusion detection is that it has the ability to analyze from audit data without the need of the admin to have obvious detail of different attack classes. Clustering can be categorized into many forms in terms of input data: Hierarchical clustering for the connectivity model, K-mean for the centroid method, distribution environment for the expectation maximization method, DBSCAN (Density-based Spatial Clustering of Application with Noise), and clique for the graph type model. Hendry and Yang [27] worked with a modified density-based method for clustering where its skills and weaknesses were discovered. The SLCT (Simple Logfile Clustering Tool) is used for the purposed approach, and it is an application used for the offline data mining tool. First, they drive that density-based clustering made the skills to minimize wide dataset nicely as well as clustering could efficiently mark suspicious activity from the normal activity by using the KDD dataset. Those properties which showed the potential of clustering can make adaptive signatures as the occurrence of attack. For the anomaly and hybrid detection area, Blowers and Williams [28] implemented the DBSCAN technique to collect normal versus anomalous attacks. The KDD dataset used for implementation and performance gave 98% for attack or nonattacked detection, and moreover the study made a good example for briefing the methods of ML for cyberoperations. Sequeira and Zaki [29] proposed a method and named it ADMIT (anomaly-based intrusion detection technique) with use of the KDD dataset which achieved 80% detection rate with a 15% FAR ratio. They compared it with the past study where the detection rate and false positive rate were 74% and 28%, respectively, on the same data.

4.6. Ensemble. Mostly, the ensemble technique is used to make learning methods strong, and these methods attain achievement by joining many learners. Adaboost [30] is a famous and popular method which is used to minimize overfitting difficulties coming from ML. The bagging or bootstrap technique is used to recover generalization of the analytical model to minimize overfitting problems, and it depends on averaging the model. The random forest method ML technique joins decision trees with ensemble methods. It is made by many trees which use arbitrarily picked data attributes for their input. The result is based on the decision of majority voting. Overfitting solution, no requirement for feature selection, and model variance minimization as the number of trees in forest increases are the important benefits of random forest. This model has few disadvantages including model complexity and lower efficiency which are due to correlated variables. Gharibian and Ghorbani [31] used the analytical method; Naïve Bayes and Gaussian used the decision tree and random forest method. The techniques detect four attacks (Dos, probe, U2r, and R2l); the accuracies noted by the random forest method are 87%, 76%, 5%, and 35%, respectively. Zhang et al. [32] provided solution to detect outliers, a signature-based threat prediction method by using the KDD dataset, and they applied the outlier's detection by the random forest technique. Their approach made patterns to determine the outliers. The modification in the original outlier detection method improves the performance and reduces the complexity of calculation. The result proves the approach and achieves a good detection rate with low false alarm. Bilge et al. [14] detected botnet detection with random forest technique by using the Netflow dataset. The paper does not include the number of trees and average of attributes; it was tested on real work network and gave TP rate 65% and FP rate 1%.

4.7. Naive Bayes. It is a simple probabilistic classifier [33] based on the Bayes theorem. It gives a way to calculate the future probability $P\ (c–x)$, from $P(c)$, $P(x)$, and $P\ (x–c)$. It assumes that the result of the value of the predictor (x) on the given class (c) is liberated than the other predictors. The theory is known as "class conditional independence." The NB classifier can control random number independent features, and they minimize high-dimensional bulk expecting the task to be single-dimensional kernel bulk estimation. One of the good points that the NB classifier has is an online method, and its training process happened in linear time.

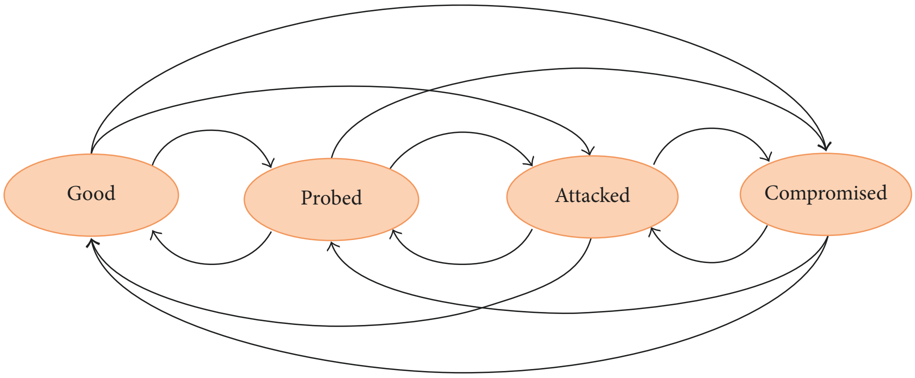


Figure 4: Hidden Markov model example.

Panda and Patra [34] worked with Naive Bayes classifier with KDD 1999 from training and test. Dataset contains probe, Dos, U2R, and R2L with the accuracy test was 96%, 99%, 90%, and 90%, respectively. After making a comparison, the NN-based method proposed the method that got better results, rather than resulting in somewhat extra false positive ratio.

For anomaly- and hybrid-based detection, Amor et al. [35] used the Naive Bayes classifier to form Bayesian network by use of the KDD 1999 dataset and categorizes into four segments to show the different attack environment and process measure. The first set includes normal data and single attack, and the second set includes all four types of threat which is presented in the KDD dataset for solution of multiclass classification. The third part contained normal data and the fourth type to sole the anomaly-detection trouble. The result is given in the paper as 97%, 96%, 9%, 12%, and 88% for normal, DoS, R2L, U2r, R2L, and probe, respectively. Due to 97% accuracy for the normal attack, FAR can be assumed to be 3%. For anomaly detection, accuracy resulted as 98% and 89% for normal and abnormal classes, respectively.

4.8. Support Vector Machine. SVM is a supervised ML algorithm to help for classification/regression problems. It based on searching an unravelling hyperplane in the feature space in two classes in such a way that the space between the hyperplane and the nearest data point of every class is maximized. This method depends upon reduced classification threat somewhat that is an ideal classification. SVM has good reputation for its generalization ability and binary classifier, and multiclassed classification is understood by making an SVM for every pair of class.

Li et al. [36] worked in the area of misuse-based detection and proposed that efficiency of IDS which is basically based on data feature dimension can be adopted by the feature removal method. They used the combination of clustering, ant colony, and the SVM well-organized classifier made for detecting whether the network traffic is normal or not. The KDD dataset used with 10-fold cross validation and 99.6% accuracy was attained in the experimental work.

Wagner et al. [37] used Netflow records and applied the support vector machine approach to detect anomalies from the network traffic. Their approach used a special kernel technique which takes into account equally the contextual and measurable information of Netflow data. The performance recorded by these criteria was from 89% to 94% in terms of accuracy for different types of attacks with FP rate from 0% to 3%.

4.9. Hidden Markov Models. HMMs are generally basics for developing probabilistic models of linear order. It gives theoretical package to build composite models that just drew an instinctive picture. For example, for the HMM-based host intrusion detection is shown in Figure 4. Here, every host has four stages which are good, probed, attacked, and compromised. If the host is in the form of indication by the source, it makes transition to the state pointed by the designation. P defines the probability between states of the model. Q observed the probability receiving different observations of the host certain state $]\pi[$ is the stating state distribution. The HMM is represented by $(P,\ Q,\ ]\pi[)$ [38].

Ariu et al. [39] proposed the HMMPayl framework to examine HTTP-payload with the HMM, and they explored new ideas during the work. They found new idea to extract features which boost the capacity of HMM to model the sequence of records. The experimental work showed a new approach that gave an analytical model of payload which is most accurate to detect the attack a with low FP rate. HMMPayl is effective for Cross Site Scripting and SQL-Injection, and they also studied the DARPA 1999 dataset with other dataset HTTP dataset.

Joshi and Phoha [40] made a study by use of HMM to build the anomaly detection framework. By using the KDDcup 1999 dataset, they showed that the proposed approach has capacity to categorize network traffic in attacked or normal. Experimental work was done on five features of KDD dataset (Src_Bytes, Dst_Bytes, duration, Is_Host_login, and Is_Guest_login), 79% accuracy was detected, and the remaining was nominated by the FP rate. The cause of the FP rate was the amount of features (12.195%) which were selected in the training period instead of all 41 features, and efficiency can be improved by tuning the HMM parameters or by using the larger dataset.

5. Analysis for Datasets

Mostly, studies used DARPA 1998, DARPA 1999, DARPA 2000, or KDD 1999 dataset, and only few used Netflow, DNS, and TCPdump data for the intrusion detection cybersecurity purpose. The reality is that many papers use the DARPA and KDD dataset to get data that are much time-consuming although if we have already had dataset and reuse which can gives use easy comparison of accuracy of different techniques. Limited use of the Netflow shows that it has no rich features like Tcpdump, KDD, and DARPA. Another issue connected to the performance of IDS is the kind of ML/DM method applied and overall system layout. The study gets concerns with many DARPA and KDD dataset and applied them with many kinds of ML algorithms. These researches did not make an IDS but inspected the performance of the ML/DM method in Internet security.

6. Tools for DM and ML

There are four free and most famous software including RStudio, RapidMiner, Pandas, StatsModels, and Scikit-learn (python libraries) and Weka. RStudio is an IDE for analytical computing and graphics. It is a core programming language known as *R* language. It has console, code highlighting editor for the direct lunching code with utility for history, debugging, and so on. RStudio has different edition including free of cost, business use, and web application platform. After *R* functionality, many high-level scripting languages like python are used by students *R* especially for sample regression and correlation analysis. RapidMiner is a good tool which provides the platform for data mining, machine learning, text mining, predictive analysis, and so on. It can create statistical workflow getting data from many data sources. RapidMiner is used for business, commercial, research, education, and training purposes. It is compatible with the stages of data mining like data preprocessing, creating model, graphical result, and so on. It is an excellent educational tool in the field of data science. Nowadays, Python is the most popular multipurpose programming language. It has packages, libraries, and framework for data analysis and visualization, which are related tasks. Pandas, StatsModels, and Scikit-learn are the three open-source libraries. It can be best to use all three libraries in IPython. "Pandas" is an open-source library that gives high performance easy to use data analysis tools for the Python language; "Statsmodel" is also the Python library which facilitates programmer to data explore, estimate analytical model, and perform statistical test. "Scikit-learn" is another open-source machine learning library for Python. It has many classifications, regression, random forest, K-means, and so on. It is a highly useful library to explore data and estimate analytical model and test. Weka is used to implement algorithms related to data mining and machine learning based on Java. It has tools for data preprocessing, regression analysis, clustering, classification methods, association rules, and so on. Weka is free to use and has user-friendly GUI.

7. Evaluation of DM/ML Methods

The mentioned study has limited literature about the evaluation of DM and ML methods according to the performance. According to the experimental comparison [41], RF, ANN, and bagged tree make the good result. After standardization, SVM and boosted tree make good performances. The study also showed that performance of specific ML methods depends on application and ways of implementation. As at the start we mentioned that the literature highlighted in this study comes from the cable-connected environment, when the system is designed to work as online application, it must handle streaming data, that is, buffering while showing the output with proper time info. Few literature explained [29, 32] their concept as processing the data online in real time. Sequeira et al. [29] used shower algorithm like sequence mining for intrusion detection, so shortly the training model involves minimum few factors including update feature, appropriate timing information, and capacity of simplification.

Update feature: Clustering method, HMM, and BN have the ability to update incrementally when compared to ANN and SVM methods which may make complications. *Appropriate timing*: Time management for each method is important to consider for streaming the model using slower methods like ANN, and sequence mining can be used for streaming if available data windowed capacity of simplification (capacity of generalization) is needed to build a strong training model. Many DM and ML have good generalization or simplification ability. The test case is mostly rapid and fast, so if the model is trained once it can be ready for online use.

8. ML and DM Usage for Misuse/Anomaly Detection

IDS is mostly based on hybrid methods and has capacity to detect anomalies and perform misuse-based detection. Anomaly detection for irregular network traffic and the misuse-based detection are based on the attacked pattern with identified signatures or expand new signatures from labeled data from anomaly modules. Mostly anomaly detection depends upon clustering, a density-based approach; for example, DBSCAN is most helpful, is easy to use, has a low parameter or is distribution based, and has a fast processing speed. For anomaly detection, SVM also performs good learning by extracting association rules from existing typical data flow. In misuse detection, signatures need to be identified, and it an important classifier that must be able to product readable signature such like branch structure decision tree, association rules, and black box classifier such as ANN SVM not suitable for misuse detection. Some approaches are statistical like Bayesian network and HMM, and some are based on entropy like decision tree. Random forest comes in the ensemble method, and some depends on upload association rules. The system designer must explore where training data have recommended quantity and have statistical qualities which can be explored. It also should be in knowledge whether the required system will be online or offline. The knowledge of such like points will determine the good available ML method. By the authors of this research, traffic data cannot be correctly modeled by using simple distribution because of the reality that, in practice, single network packet may contain payload which can be associated with a number of network protocols/user behaviors. The unevenness in payload is considered by collection of more than one probability distribution which is not openly divisible. So in algorithms like Bayesian network, the HMM might not have a suitable approach due to data having no properties that are suitable for them. If training data are unusual, the random forest may have benefits; if attack signature collection is significant, decision tree and association rules can become helpful.

9. Evolving Problems and Probable Solution

The problem with so many methods surveyed is that each of the techniques generate many false alarms. Low detection of R2L and U2R is also a problem which can hap due to these attacks similar to normal data and so many times misclassified. Another reason for this problem is less regular occurrence of these attack classes which cause classifier biasness and reduced detection rate .The experimental work mostly done on benchmark dataset training and test in two different environments reduced performance. IDS alarm has too many system admins to deal for every attack freely. The strategy to develop group alarm and raising alarm for every group must be developed. All methods need much training and testing time which is unwanted, so no techniques are flexible for real environment. There are many research studies available for anomaly-based detection, but still mostly, IDS has misuse detection due to model plans all based on labeled data, the point which we do not have in real network flow.

9.1. Probable Solution. To avoid false alarm ratio, IDS must have online learning capability, drift concept handling, and capacity to adjust in any environment. To enhance the detection ratio for R2L, U2R feature selection, clustering, classification techniques, and selection of attributes which are very specific for two classes should be chosen. Biasness problem can be reduced by conducting skewed class delivery. Training and test time can be minimized to avoid the usage of features of data denoted as a single point and transformed as single-dimensional data for training and test.

10. Conclusion

This study demonstrates the literature survey-based on machine learning/data mining techniques for Internet/communication security. The study highlighted the papers that define the use of multiple ML/DM methods for Internet domain for both misuse and anomaly-based detection. Unfortunately, as the techniques which are more reliable and accurate for mentioning the domain are still undiscovered, the present richness and complexity of techniques are not possible to recommend for one single method, based on the cast of attack the system supposed to diagnose. ML and DM techniques do not work without symbolic data, and it is hard to create these kinds of datasets. The largest break noticed is the presence of labeled data, and it is understood to need useful deal to gather data and label it. For usage of new data, some intelligent DM and ML techniques can develop models, comparison, and reduce the list of DM and ML uses for Internet applications. Major advancements can be made to ML and DM techniques in communication or Internet security by using this dataset, and discoveries can be conceivable. The rich area of the study is to explore the techniques of fastest incremental learning which can be useful for day to day update of models for misuse and anomaly detection.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, “Network anomaly detection: methods, systems and tools,” *IEEE Communications Surveys and Tutorials*, vol. 16, no. 1, pp. 303–336, 2014.

[2] N. Padhy, P. Mishra, and R. Panigrahi, “The survey of data mining applications and feature scope,” *International Journal of Computer Science, Engineering and Information Technology*, vol. 2, no. 3, pp. 43–58, 2012.

[3] S. Agrawal and J. Agrawal, “Survey on anomaly detection using data mining techniques,” *Procedia Computer Science*, vol. 60, pp. 708–713, 2015.

[4] M. Ahmed, A. N. Mahmood, and J. Hu, “A survey of network anomaly detection techniques,” *Journal of Network and Computer Applications*, vol. 60, pp. 19–31, 2016.

[5] R. F. Najeeb and B. N. Dhannoon, “Classification for intrusion detection with different feature selection methods: a survey (2014–2016),” *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 7, no. 5, 2017.

[6] S. Fong and H. Yang, “The six technical gaps between intelligent applications and real-time data mining: a critical review,” *Journal of Emerging Technologies in Web Intelligence*, vol. 3, no. 2, pp. 63–73, 2011.

[7] A. Guazzelli, M. Zeller, W. Chen, and G. Williams, “PMML an open standard for sharing models,” *R Journal*, vol. 1, no. 1, pp. 60–65, 2009.

[8] C. Shearer, “The CRISP-DM model: the new blueprint for data mining,” *Journal of Data Warehousing*, vol. 5, pp. 13–22, 2000.

[9] H. Altwaijry and S. Algarny, “Bayesian based intrusion detection system,” *Journal of King Saud University–Computer and Information Sciences*, vol. 24, no. 1, pp. 1–6, 2012.

[10] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, “A detailed analysis of the KDD CUP 99 data set,” in *Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defense Applications*, Ottawa, ON, Canada, July 2009.

[11] C. Thomas, V. Sharma, and N. Balakrishnan, “Usefulness of DARPA dataset for intrusion detection system evaluation,” in *Proceedings of the Data Mining, Intrusion Detection, Information Assurance and Data Networks Security*, Orlando, FL, USA, March 2008.

[12] J. P. Gifty, M. Ravichandran, and C. S. Ravichandran, "Efficient classier for R2L and U2R attacks," *International Journal of Computer Applications*, vol. 45, no. 21, 2012.

[13] C. Kruegel and T. Toth, "Using decision trees to improve signature based intrusion detection," in *Proceedings of the 6th International Workshop Recent Advances Intrusion Detection*, pp. 173–191, West Lafayette, IN, USA, 2008.

[14] L. Bilge, S. Sen, D. Balzarotti, E. Kirda, and C. Kruegel, "Exposure: a passive DNS analysis service to detect and report malicious domains," *ACM Transactions on Information and System Security*, vol. 16, no. 4, pp. 1–28, 2014.

[15] S. Aljawarneh, M. B. Yassein, and M Aljundi, "An enhanced J48 classification algorithm for the anomaly intrusion detection systems," *Cluster Computing*, pp. 1–17, 2017.

[16] B. Ahmad, W. Jian, B. Hassan, and S. Rehmatullah, "Hybrid intrusion detection method to increase anomaly detection by using data mining techniques," *International Journal of Database Theory and Application*, vol. 9, no. 12, pp. 231–240, 2016.

[17] B. Ahmad, W. Jian, and M. Shafiq, "Intrusion detection by using hybrid of decision tree and K-nearest neighbor," *International Journal of Hybrid Information Technology*, vol. 9, no. 12, pp. 201–208, 2016.

[18] L. P. Dias, J. J. F. Cerqueira, K. D. R. Assis, and R. C. Almeida, "Using artificial neural network in intrusion detection systems to computer networks," in *Proceedings of the Computer Science and Electronic Engineering (CEEC)*, Colchester, UK, September 2017.

[19] A. S. Sodiya, O. A. Ojesanmi, A. Akinola, and O. Aborisade, "Neural network based intrusion detection systems," *International Journal of Computer Applications*, vol. 106, no. 18, pp. 19–24, 2014.

[20] D. P. Vinchurkar and A. Reshamwala, "A review of intrusion detection system using neural network and machine learning technique," *International Journal of Engineering Science and Innovative Technology*, vol. 1, no. 2, 2012.

[21] H. Brahmi, B. Imen, and B. Sadok, "OMC-IDS: at the crossroads of OLAP mining and intrusion detection," in *Advances in Knowledge Discovery and Data Mining*, pp. 13–24, Springer, New York, NY, USA, 2012.

[22] K. Nalavade and B. B. Meshram, "Mining association rules to evade network intrusion in network audit data," *International Journal of Advanced Computer Research*, vol. 4, no. 2, 2014.

[23] Livadas, R. Walsh, D. Lapsley, and W. Strayer, "Using machine learning techniques to identify botnet trac," in *Proceedings of the 31st Conference on Local Computer Networks*, pp. 967–974, Tampa, FL, USA, November 2006.

[24] N. Devarakonda, S. Pamidi, V. Kumari, and A. Govardhan, "Intrusion detection system using Bayesian network and hidden Markov model," *Procedia Technology*, vol. 4, pp. 506–514, 2012.

[25] I. Alocious, C. Abouzakhar, N. Xiao, and Christianson, "Intrusion detection system using Bayesian network modeling," in *Proceedings of the 13th European Conference on Information Warfare and Security ECCWS 2014*, pp. 223–232, Piraeus, Greece, July 2014.

[26] 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," in *Proceedings of the SoutheastCon 2016*, pp. 1–6, Norfolk, UK, March–April 2016.

[27] R. Hendry and S. J. Yang, "Intrusion signature creation via clustering anomalies," in *Proceedings of the Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security*, Orlando, FL, USA, March 2008.

[28] M. Blowers and J. Williams, "Machine learning applied to cyber operations," in *Network Science and Cybersecurity*, pp. 55–175, Springer, New York, NY, USA, 2014.

[29] K. Sequeira and M. Zaki, "ADMIT: anomaly based data mining for intrusions," in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 386–395, Edmonton, AB, Canada, July 2002.

[30] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," in *Proceedings of the 13th International Conference on Machine Learning*, vol. 96, pp. 148–156, Bari, Italy, July 1996.

[31] F. Gharibian and A. Ghorbani, "Comparative study of supervised machine learning techniques for intrusion detection," in *Proceedings of the 5th Annual Conference on Communication Networks and Services Research*, pp. 350–358, Gazimagusa, North Cyprus, May 2007.

[32] J. Zhang, M. Zulkernine, and A. Haque, "Random forests-based network intrusion detection systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 5, pp. 649–659, 2008.

[33] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Mateo, CA, USA, 3rd edition, 2011.

[34] M. Panda and M. R. Patra, "Network intrusion detection using Nave Bayes," *International Journal of Computer Science and Network Security*, vol. 7, no. 12, pp. 258–263, 2007.

[35] N. B. Amor, S. Benferhat, and Z. Elouedi, "Naive Bayesian networks in intrusion detection systems," in *Proceedings of the ACM Symposium on Applied Computing*, pp. 420–424, Nicosia, Cyprus, March 2004.

[36] Y. Li, J. Xia, S. Zhang, J. Yan, X. Ai, and K. Dai, "An efficient intrusion detection system based on support vector machines and gradually feature removal method," *Expert Systems with Applications*, vol. 39, no. 1, pp. 424–430, 2012.

[37] C. Wagner, F. Jrme, and E. Thomas, "Machine learning approach for IP-flow record anomaly detection," in *Networking 2011*, pp. 28–39, Springer, NewYork, NY, USA, 2011.

[38] A. Arnes, F. Valeur, G. Vigna, and R. A. Kemmerer, "Using hidden Markov models to evaluate the risks of intrusions: system architecture and model validation," in *Lecture Notes in Computer Science*, pp. 145–164, Springer, NewYork, NY, USA, 2006.

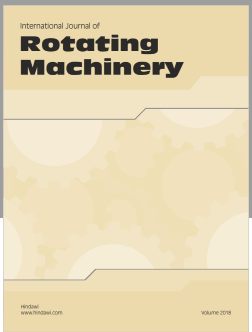
[39] D. Ariu, R. Tronci, and G. Giacinto, "HMM-Payl: an intrusion detection system based on hidden Markov models," *Computers & Security*, vol. 30, no. 4, pp. 221–241, 2011.

[40] S. S. Joshi and V. V. Phoha, "Investigating hidden Markov models capabilities in anomaly detection," in *Proceedings of the ACM 43rd Annual Southeast Regional Conference*, vol. 1, pp. 98–103, Kennesaw, GA, USA, 2003.

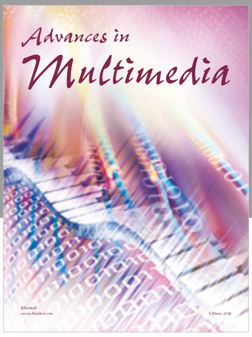
[41] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the ACM 23rd International Conference on Machine Learning*, pp. 161–168, Pittsburgh, PA, USA, June 2006.

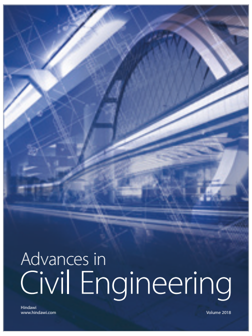
Journal of
Engineering

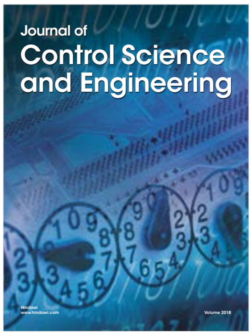
The Scientific World Journal


International Journal of
Rotating Machinery

Journal of
Sensors

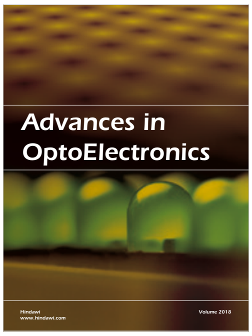

Advances in
Multimedia


Advances in
Civil Engineering


Journal of
Control Science and Engineering

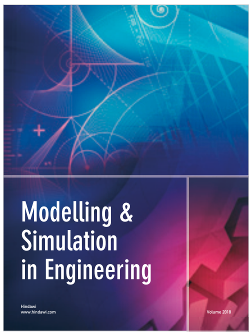
Journal of
Robotics

Journal of
Electrical and Computer Engineering

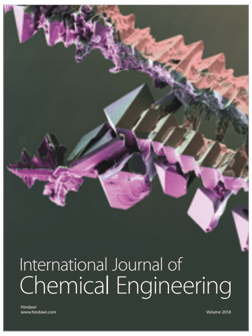

Advances in
OptoElectronics

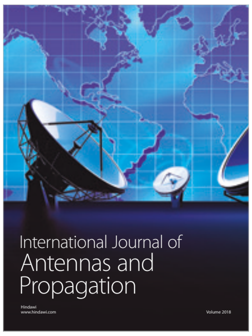
VLSI Design

International Journal of
Navigation and Observation

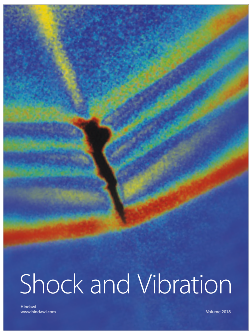

Modelling & Simulation in Engineering

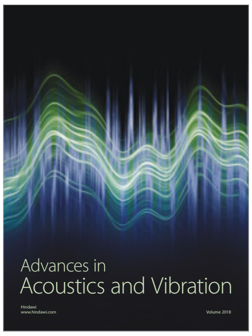
International Journal of
Aerospace Engineering


International Journal of
Chemical Engineering


International Journal of
Antennas and Propagation

Active and Passive Electronic Components


Shock and Vibration


Advances in
Acoustics and Vibration