IOSR Journal of Engineering (IOSRJEN) www.iosrjen.org
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 08, Issue 8 (August. 2018), || V (V) || 25-42

A Comparative Survey on the Influence of Machine Learning Techniques on Intrusion Detection System (IDS)

B. Narendra Kumar[1], Dr. M S V SivaramaBhadri Raju[2],
Dr. B. Vishnu Vardhan[3]
[1]*Associate Professor, Department of CSE, SSJ Engineering College, Hyderabad, Telengana, INDIA.*
[2] *Professor, Department of CSE, SRKR Engineering College, Bhimavaram, AP, INDIA.*
[3]*Professor, Department of CSE, JNTUCEM, Peddapalli, Telangana, INDIA*
Corresponding Author: B. Narendra Kumar[1]

Abstract--With the enormous growth of computer networks and content usage of users, there is a need for secure and reliable networks. As it is observed that the different types of network attacks are raised over a period of time, it is necessary to make the availability of effective automatic tools in order to identify the attack detection scenarios. Intrusion Detection System is one of the attack detection systems that detect intrusions coming from the Internet. Several approaches were observed in the literature for intrusion detection over the network. In the recent past, mining techniques were prevalent in order to check the intrusion detection. The characteristics of incoming intrusions were identified by using the mined knowledge over the data present in the network. Whenever a matching is found in the characteristics of the mined data then it is declared as an intrusion. Based on this criterion various intrusion detection models were developed in the recent research and the accuracy is improved. In this paper, a brief review is carried out over the earlier approaches. The complete approaches are divided into data preprocessing approaches and detection approaches. Further, the data preprocessing approaches are divided into Feature extraction and feature transformation models based on working methodology over the features. Similarly, the detection approaches are categorized as machine learning and evolutionary approaches. The complete details about the advantages and disadvantages of all the mentioned approaches are also described in this paper. A comparative analysis is also carried out between the approaches based on their working methodology.

Keywords-- Intrusion Detection System, Feature Extraction, Classification, Machine Learning.

Date of Submission: 16-08-2018 Date of acceptance: 03-09-2018

I. OVERVIEW

With the rapid growth in the technologies, computers and networks are under threat from worms, viruses and attacks. The number of devices connected to Internet is increased rapidly year by year. This tends to achieve about 50 billion devices by 2050 [34]. Since there is an advancement in the number of intrusions, protection of these interconnected devices and also the data passing through them is a challenging task. To address this issue, a large number of discussions against network attacks were presented in the literature. Despite all the efforts made by the researchers in the community over the last two decades, the network security problem is not completely solved. One reason for that is the rapid growth in computational power and available resources to attackers, which enables them to launch complex attacks [35]. This is considered as a two player game, where an attacker attempts to find the most effective strategy to disrupt normal operations in a network and the defender's challenge is to determine optimal defensive solutions and block illegitimate access to the network.

Generally, the defence against network attacks accomplishes in three phases-preparation, detection and reaction. In general, a security engineer conducts a risk analysis process during the preparation phase to obtain the knowledge about the environment and the data that are needed to be protected in such environment. This process is very important because it provides the sufficient information about the attacks and the effect of attacks on the network [23]. The preparation phase also includes identification of infrastructure vulnerabilities, development of security strategies and plans and installation of required security devices based upon analysis of the information gathered [19], [20]. Another key element of network security is a detection system. An intrusion detection system (IDS) [7] usually complements a firewall to form an effective cyber security solution. Fast detection of attacks is required to be able to react rapidly. Thus, an automatic detection phase is of paramount importance. Finally, handling detected intrusions in a network is carried out during the reaction phase. A traffic blocking method is an example of a mitigation mechanism used in the reaction phase. One of the main

challenges in securing networks is the appropriate design of IDSthatmonitors network traffic and also identifies network intrusions, effectively.

II. INTRUSION DETECTION SYSTEM

IDS [7] is a sort of security administration framework for computers and networks. An IDS accumulates and examines data from different regions inside a computer or a network to distinguish conceivable security breaches, which incorporate both intrusions (attacks from outside the organization) and misuse (attacks from inside the organization). The general model of a network communicating system with IDS is shown in Figure.1.

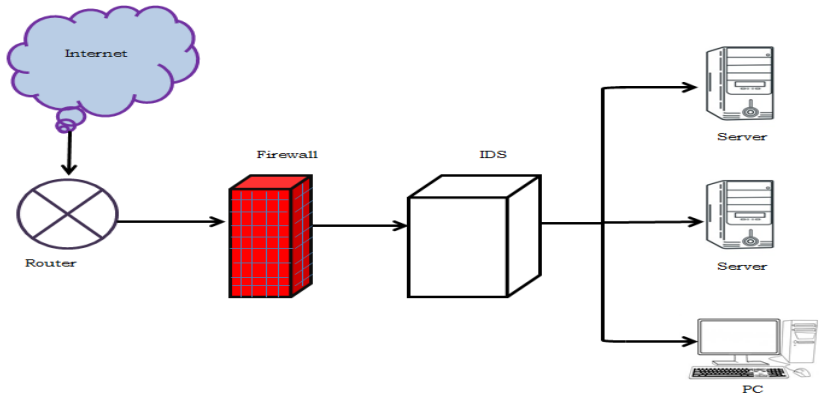


Figure.1. General Communication system throughIDS

In the above General Communication System, the IDS identifies the Internet traffic with abnormal characteristics to search either for characteristics of known traffic or deviations of normal activity. An IDS is a dynamic checking element that supplements the static observing capacities of a firewall. An IDS screens traffic in a network in promiscuous mode, especially like a network sniffer. The network packets that are gathered for rule violations are detected by a detection algorithm. At the point when rule violations are recognized, the alarm of the IDS bells. An IDS is fit for distinguishing a wide range of malicious network traffic and computer usage. This incorporates network attacks against vulnerable services, data driven attacks on applications, for example, benefit heightening, unauthorized logins and access to touchy records and malware. The main functionalities of IDS include the following.

1. Observing and analyzing both user and system activities.
2. Capacity to recognize patterns typical of attacks.
3. Investigation of abnormal activity patterns.
4. Tracking user policy violations.
5. Analyzing system configurations and vulnerabilities.

2.1 Host and Network based IDS

Considering the source of data being used for intrusion detection, IDSs are classified into two categories- Host-based IDS (HIDS) and Network-based IDS (NIDS) [12]. A HIDS operates on a single host and monitors eventsoccurring within an individual computer system. So, HIDS provide protection for critical computers that may house sensitive information. On the other hand, NIDSare not restricted to packets going to a specific host since all the machines in the networkare protected using this NIDS. A NIDS monitors traffic in a network segmentand analyses the traffic in order to identify suspicious activities. A comparative analysis [9] between the HIDS and NIDS is represented in table.1.

Table.1. Comparative analysis between HIDS and NIDS

Performance Criteria	**Host Based IDS**	**Network Based IDs**
Intruder Detection	Inside intruders are detected strongly	Outside intruders are detected strongly
Intruder Prevention	Inside intruders are prevented strongly	Outside intruders are prevented strongly
Response time	Response Time is short for real time aspects. Works effectively for long term attacks	Response Time is long for outside intruders
Response to damage	Excellent in determining the damage extent	Poor performance in damage detection.

2.2 Signature and Anomaly Based IDS

It is also possible to classify the IDS through the methodology used for detection:signature basedintrusion detection [42] and anomaly based intrusion detection [31]. A signature detection system distinguishes patterns of traffic or application data expected to be malicious while anomaly detection systems compare activities against a normal pattern.

2.2.1 Signature Based IDS

Signature based IDS [42] includes scanning network traffic for a progression of bytes or packet sequences known to be malicious. The essential preferred standpoint of signature detection is that known attacks are detected reliably with a low false positive rate. A key advantage of this detection method is that signatures are easy to develop and recognize if there is information about the network behavior. The major drawback of the signature detection approach is that such systems ordinarily require a signature to be characterized for all of the possible attacks launched by an attacker against a network. Signature based detection systems are likewise inclined to false positives since they are commonly based on regular expressions and string matching. Boththese mechanisms only search for strings within packets transmitting over the wire. While signatures work well against attacks with a settled behavioral pattern, they do not work well against the multitude of attack patterns created by a human or a worm with self-modifying behavioral characteristics. Detection is further complicated by advancing exploit technology that permits malicious users to conceal their attacks behind payload encoders and encrypted data channels.

2.2.2 Anomaly Based IDS

The Anomaly Based IDS [31] centers on the concept of a baseline for network behavior. This baseline is a description of accepted network behavior, which is learned or specified by the network administrators, or both. Events in an anomaly detection engine are caused by any behaviors that fall outside the predefined or accepted model of behavior. Anomaly detection systems have two major advantages over signature based IDSs. The first advantage that differentiates anomaly detection systems from signature detection systems is their capacity to detect unknown attacks. This advantage is because of the capacity of anomaly detection systems to model the normal operation of a system/network and detect deviations from them. The second advantage of anomaly detection systems is that the previously mentioned profiles of normal activity are modified for each system, application and/or network, and therefore making it very difficult for an attacker to know with certainty what activities they carry out without getting detected. In any case, the anomaly detection approach has its share of drawbacks too. For example, the inherent complexity of the system, the high percentage of false alarms and the associated difficulty of determining which specific event triggered those alarms are some of the many technical challenges that need to be addressed before anomaly detection systems can be widely adopted.

Based on these aspects various intrusion detection approaches are proposed in earlier and a brief survey is carried out over all the earlier approaches, which aredescribedinthenextsection.

Table.2 Signature based IDS vs. Anomaly based IDS

	Signature Based IDS	**Anomaly Based IDS**
Pros	1. Simple and effective. 2. Effective in the known attack detection.	1. Effective in new and unknown attack detection. 2. Facilitate detection of privilege abuse.
Cons	1. Ineffective in unknown attack detection. 2. Hard to update the signatures. 3. More time consumption for a large dataset.	1. Weak accuracy profile due to random changes. 2. Difficult to trigger alarms in right time. 3.High False Positive Rates.

III. LITERATURE SURVEY

This section describes various approaches proposed earlier to perform anomaly based intrusion detection. The basic block diagram for the anomaly basedIDS is shown in Figure.2. It shows the basic functional block diagram of an anomaly based IDS. Here the complete system is accomplished in two phases, TrainingandTesting.

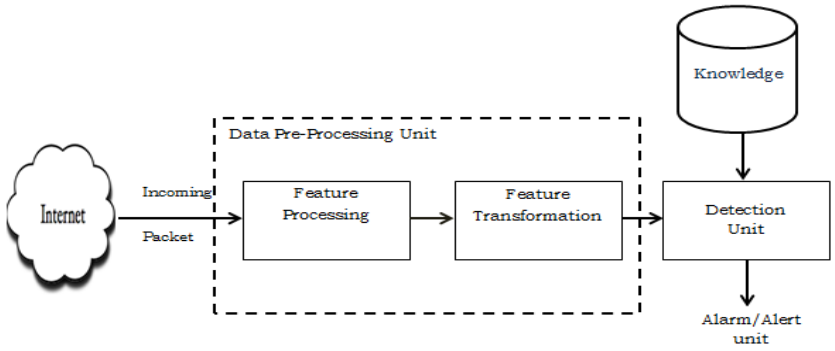


Figure.2 Simple block diagram of anomaly based IDS

Training phase performs the knowledge creation about the characteristics of network traffic. The Testing phase involves the detection of an intrusion. In this detection system, the detection unit detects the intrusions based on the features of incoming packets. In this scenario, features of packets are very important to obtain maximized detection accuracy. Along with this accuracy, the effective feature processing also reduces the computational overhead. The combined unit of Feature Processing (FP) and Feature Transformation (FT)iscalled as a Data Preprocessing Unit (DPU).

3.1 Data Preprocessing

The function of Data Preprocessing unit (DPU) is to convert the network traffic into different sequences. These sequences primarily reflect observations. Each observation has a character of its own and it is noticed as a representation of a feature vector. Further these feature vectors are optionally labeled as "Normal" and "Anomalous". The main objective of a DPU is to convert the network traffic such that it is suitable for data mining or machine learning algorithms. Most of the approaches considered the packet header information as processing features.

3.1. Feature Extraction

The process of creating features for a given learning or classification instance is called feature extraction**.** The Feature Extraction depends on packet header information**.** The packet header carries a very little segment of data from the complete network traffic. So, the processing header data requires lesser sources (memory, storage, and processing unit) than processing the entire packet data. Anomaly detection is a very important process. If this process is carried out depending only on the information of packet header results in the reduction of data preprocessing requirements of data.This review considers three type of packet header information. They are the Basic Features, Time Based Statistical Features (TSF) and Time Based Statistical Multiple Features (TSMF) [24] features.Ontu et.al., [24] RecognizedBasic,TSF and TSMF as the prime categories. They continued to sub classify the mentioned. Itproducesafine grained graph. This graph represents 26 feature categories.

3.1.1.1 Basic Features

The Basic Features are generally extracted from the packet header to detect the attacks against wireless networks. Theframe header of MAC layer is considered for analysis in the given scenario. This process requires a local Wireless Network to be tapped. Guennoun et.al, [29] processed for feature extraction and new feature derivation from the obtained features by extracting all the frame headers and converting them into any continuous features to categorical features. A feature selection approach is designed for the detection of similar features. This is intended to detect malicious traffic. The duty of a filter is to assess the information gain ratio. This is performingindividually. The outcome of the above step is the retrieval of several features sorted by relevance. Then a wrapper approach is applied over the list of features to obtain best feature set. The most popular forward search algorithm was accomplished to search the relevant feature starting with single most relevant feature: k-means classifier is used for testing and then the relevant next most features are added iteratively to the set. This approach declaredthat the top eight ranked features produce best classifier accuracy.

Stealthy scans are detected byStatistical packet anomaly detection engine (SPADE). SPADE is an anomaly detection engine. This works on Statistical packets.SPADE is proposed by Stanford et.al, [13]. SPADEconsidered the basic features rather than constructing model. This is atraffic distribution model scanning the networks which are observed using Bayesian Networks or Joint Probability Measurements.

According to Early et.al, [22], packet header features never be blindly used. This results ininaccurate classification. Most of the headers do not have inherent anomalous value. As a result of this they are generally irrelevant and it is also not feasible for the complete exercise of these values. The nature of this experiment

appears to contradict the approach of Packet header anomaly detector (PHAD) [10]. Allthe basic features along with few irrelevant ones are used in this process.PHAD forms clusters and mitigates accuracy programs. Clustering reduces false positives, by ensuring unnoticed legitimate values. SPADE removes all the irrelevant featureswith the help of tiny subset. This subsetis made of packet headers. IDS by Guennoun et al. [29] is based on used feature selection techniques intended for the elimination of such unrelated features.

3.1.1.2 Time based statistical features (TSF)

TSF features aregenerated by observing the basic features. This monitoring is performed over the time of such flow. Examples coverpackets along with bytes. The inter packet arrival time andthe mean packet length are also included in the examples. These features play an important role in the sessions of fingerprinting. They help in the detection of unusualdata flows. Further,they are also useful in identifying other anomalies in one session.

Ramdas et.al, [15] proposed an IDS approach using the time based statistical features. Here the considered TSF features are Start time, End time, Quad4, whether there is valid start of session, whether the connection was closed properly or improperly, average size of questions, average size of answers, number of queries per second, idle time between question and answer, ideal time between answer and questionand the duration of connection. Self-organization Map was used at detection phase to compare the data instances for the anomaly detection in that particular service.

TSF features are also used by Early et.al, [22]. Their main intention is to find the application protocol automatically spearing the use of destination port as guide. This approach considers the features form only TCP/IP packet headers. In this approach C5 decision tree is used as a classifier. The main features aremean packet length and mean packet inter-arrival time. The percentage of packets is set. This is arranged with TCP state flags. The anomaly detection mode makes use of the mentioned method for the detection of working with the ports which is not standard. These are possibly flagging backdoors.

Yamada et al, [25] used TSF features for thedetection of attacks on web servers in the condition of encryption. The featuresconsidered in this approach are request size and response size of HTTP, measured along every regular activity of user. If size attributes are used alone, they create a different scenario andsuch usage results in the yield of large number of false positives and the performance of frequency analysis. Such performance is intended for the removal of alerts which are very general to web server. In terms of statistics,uncommonalertsare the anomalies.

TSF features are used for the detection of links which go along many steps. This is as mentioned by Yang and Huang et.al, [26]. The assumption is that such links are used to prevent for being followed by attackers. The time taken for the round-trip times (RTTs) is calculated. Based on this calculation depends detection. This whole thing is connected to the packets in TCP connection. The above method makes use of clustering and partitioning algorithm for calculation of RTTs. It is also used for the calculation of number of stepping units. Only the packet header information is used in this approach, particularly the timestamps of the Send and Receives packets.

TCP flags are used to build the anomaly detectors for TSF features [16], [32]. TCP flags collected from each TCP session packets, and the combination of each flag is a quantized symbol. Thus, the TCP session is converted into a sequence of symbols and modeled through markov chain model. For every observed protocol, a separate model is generated. Traffic in given network is evaluated. This is carried out against the predefined models for anomalous detection in detection phase.

TSF features are beneficial in the identification of behavior which is anomalous in singlesession, like"an unexpected protocol, unusual data sizes, unusual packet timing, or unusual TCP flag sequences". However, they have few limitations in finding the activity covering multiple flows. The example of such flow is Denial of Service attacks i.e., DOS.For that, the TSFfeatures for multiple flows are required. These are called Time Based Statistical Multiple Features (TSMF).

3.1.1.3 Time based Statistical Multiple Features (TSMF)

For the consumption of TSMF features, base features are used. This is performed by observing their flow over multiple connections. They are found to be better in exhibiting discrimination among traffic patterns. Thisincludes the comparison between normal and anomalous patterns basic features.

Dickerson et.al, [8] proposed an anomaly basedIDS based on the TSMF features and the fuzzy logic and it is named as "Fuzzy Intrusion Recognition Engine (FIRE)". TCP flags, quad and the length of packet are the features extracted from the network traffic. From these basic features, the TSMF features evaluated through some statistical measures as: "the number of new source-destination pairs seen, and the number of new source-destination pairs which are not in the long termdatabase". Fuzzy threat analysis functions effectively when takes an input. Every TSMF feature performs this action of preparing the input.

Anomaly detection technique is a very important technique with significant contribution. It is proposed by Barbara et.al, [11]. "Audit data analysis and mining (ADAM)", derived the TSMF features through the association mining rules. In this approach, the DPU takes up the responsibility of tracking the flows in the networks. Further it creates the connection records. These records have the below mentioned primary qualities. This primary quality is also called as Basic Features. First of these features is quad. This is followed by start time along with connection status. Association mining is used to execute several tasks. Such tasks are carried out by applying this to connection records through one floating window. The size of the window is very important in defining certain features. It defines various patterns which are looked as top support values. During the trainingphase, a model is generated to signify a type of behavior. This model is based upon association rules. The entire process is intended to signify a normal systembehavior. In thedetection phase association rules are found. Data mining finds these rules compared to anomaly detection models.

Ertozet.al, [18] proposed an IDS, "Minnesota IDS (MINDS)", is a flow based approach considered the flow count of features. Then these statistics are given to the anomaly detection algorithm. MINDS extract the TSF features such as number of packets, number of bytes, protocol, quad and union of TCP flags by processing ten minute batches of NetFlow records. Then there TSF features are observed over a time window to derive various TSMF features. Then they are given as inputs for the "density based outlier detection algorithm named as Local Outlier Factor (LOF)" to detect anomalies. This LOF approach is compared with the conventional SVM and nearest neighbor approach to unsupervised network anomaly detection.

A new network anomaly detection approaches are proposed Lazarevic et.al, [17]. Data preprocessing even though similar tosome other form it makes use of, TCPTRACE [23]outputs ignoring NetFlow records. Similar to NetFlow, TCPTRACE analyzes packet headers only. However, it analyses thelinkswhich are in two directions instead of one-way flows. Even though the payloads of packets are not touched for the extraction of the features there are some changes to be observed. In the detection phase, few attacks are detected. One of these attacks is user to root (U2R). The other one is remote to local (R2L). However, the main disadvantage with this approach is high false positive factor which is not adaptable for its utilization. The authors declared that the TSMF based features are very important in detecting the DOS and probe attacks and the TCPTRACE basic features play a significant role in the detection of U2R and R2L attacks. Subsequently there was an extension of LOF algorithm to assist certain parameters. This is in a border prospective is, applied incremental updates. This has been mentioned in their work by Pokrajac et.al, [27].

NetFlow records are very important in the anomaly detection of networks. Quadand time fields help in developing such detectors. This is known as network anomaly detector having its genesis fromTSMF features. This Entire information is mentioned by Lakhina et.al, [21] in their research work. It is assumed that the traffic anomalies change the distribution directions of the selected header fields. Detector uses entropy measures in basic properties for five minutes to detect anomalies. Network scans and worm behavior are some of the anomalies detectable out of this approach. Besides these outages along with point-to-multipoint-traffic, Port scans, DOS attacks, alpha flows5, flash crowd6 outages are some of the anomalies that are detectable by this approach. The anomaly detector makes use of packet sampling for confirming few operations. These operations are real-time in their nature. They are worked out on highbandwidth backbone networks.

NetFlow has ability to perform 1 in N packet sampling. Here N is configurable.Lakhinaet.al, [21] carried out their work in an extensive way by keeping in view various parameters. They used sample 1 in 100 packets. This was executed while locating network anomalies. It is an intuition that, sampling packets are responsible for the reduction of the exactness of detection. This is with reference to network intrusion detection system. Hence Patchaet.al, [28]made use of adaptive sampling technique. This application was to achieve balance between accuracy need and few other important quantifiers. It is a very important phase where this is very specific task include resource overheads.

"Stochastic Clustering Algorithm for Anomaly Detection (SCAN)" Patchaet.al, [28] aims to find network anomalies even in the absence of complete and accurate audit data. SCAN approachperforms both sampling of incoming data and also creates the summaries of data to minimize the workload. The Basic Features considered are connection status, quad, protocol and duration extracted from every connection. Then by using a window of time 60 seconds, the TSMF features are calculated from these basic header features to summarize the data including- "percentage of data packets, percentage of control packets, flow concentration factor and the maximum number of flows for a particular service". Time-based properties are used by a clustering algorithm to detect anomalies. When tested, SCAN was able to detect network-based DOS attacks in high-speed networks even when data sampling was carried out.

Lu and Ghorbanifar et.al, [33] use signal processing techniques to detect anomalous traffic in theDARPA 99 dataset. The 15 custom TSMF features measured flow counts, bytes per flow, bytes per packet and packets per flow, all over a window of time period of one minute. These features were used tocreate a model of normal traffic based on the Wavelet Transform Analysis.

IPFIX data also has a role to play in this process. It is used as input for the detection of anomalies. Such anomaly detection system is mentioned by Muraleedharanet.al, [36] in their work. IPFIX is the outcome of IETF work. The objective of IPFIX is to standardize the NetFlow. This approach configured the IDS to monitor ICMP traffic and TCP traffic. It is also applied to UDP traffic to derive a record after every time window. Here the derived TSMF features are: "number of packets, average packet size, average flow duration, number of flows, average packets per flow, and number of single packet flows". These properties were used in many ways. Generating profiles of normal traffic is the most important application. It is followed a chi-square measure. The intension of the entire process is the detection of anomalies during the detection phase. This approach is competent in the detection of flood, Denial of Service(DOS), scan and Distributed Denial of Service(DDOS)attacks.

Byusing theTSMF features it is known that all the approaches were suitablein thedetection of the responses of DOS. They also suit the detection of network scan. These approaches have their limitations. They are not effective in the detection of single packet and single flow. Further, their limitations extendin the detection ofpayload based attacks as well.

Table.3. Comparative analysis of various Data Preprocessing techniques

Methodology used	**Listed Approaches**	**Advantages**	**Disadvantages**
Basic Features	Guennoun et.al. [29] Stanford et.al. [13] Early et.al. [22] Mahoney et.al. [10]	Very simple to implement. Probe, DOS attacks detection.	Lower detection accuracy.
TSF	Ramadas et al. [15] Yamada et.al. [25] Yang et.al. [26] Broadely et.al. [22]	Effective in detecting anomalous behavior in unexpected cases.	Not able to detect DOS attacks.
TSMF	Dickerson et.al [8] Barbara et al. [11] Ertoz et. al. [18] Lazarevic et al. [17] Pokrajac et al. [27] Lakhina et al. [21] Patcha et.al. [28] Ghorbanifar et.al. [33] Muraleedharanet.al. [36]	Suits the detection ofnetworks. Suits the detection of DOS behavior.	Single packet attacks are not detected. single flow attacks are not detected. Payload-based attacks are not detected.

3.1.2 Feature Transformation

Further, the obtained features are processed to transformations. The main objective of feature transformation is to reduce the extra computational overhead of the detection system. The overall size of feature set is considerably high which inturn consumes more time and also more computational iterations. Hence the size of the feature needs to be reduced. Previously various approaches are developed to perform feature transformation. All the Feature Transformation Approaches are classified as the Feature Reduction Approaches and Feature Selection Approaches

3.1.2.1 Feature Reduction

Feature Reduction is a concept in which fresh subspacesare found withfewerdimensions. This comparisonis with original feature space [30]. The popular methods for feature reduction are "Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), Uncorrelated Linear Discriminant Analysis (ULDA), GLDA [67] and Independent component analysis (ICA)".

Wahba et.al, [63] proposed a multiclass classification to help in creating present day IDS models with an improved accuracy and efficiency. The objective is to join different classifiers to accomplish better outcomes. The authors emphasize the significant benefits of utilizing multi-layered IDS supported by recent research investigations and methodologies. The author also determines that the overall performance of the model is impaired during the application phase. This is because of the merging of all classification patterns where there is no possibility of feature match, this results in data redundancy [63]. Another approach is proposed by the author to reduce the irrelevant features,thereby increasing the performance of system.

Tesfahunet.al, [44] disclose that because of the unstable nature of IDS, there exists a greatdiscrepancy in "NSL-KDD (National Scientific Laboratory-Knowledge Discovery and Data Mining)" dataset.Class imbalances have to be carefully addressed to achieve efficiency. In this process, synthetic minority sampling

technique has a great role. This technique is used over training data. The feature selection method based on information acquisition is used to create a subset of reduced features. Due to the presence of the large number of redundant records in NSL-KDD dataset, it is chosen as a benchmark dataset, avoiding the detection of minority classes such as R2L and U2R. To test the proposed approach, NSL-KDD data set was used and the number of features has been brought down to 22, the U2R detection rate is increased followed by a reduced build time.

Desaleet.al, [65] proposed an approach for effective IDS.This approach is based on Genetic Algorithm. The genetic algorithm has a great role to playin the exploration ofmethod. This is performed during feature selection from complete NSL-KDD data set. The obtain results prove the advantages of the proposed approach "Mathematical Intersection Principle". This is applied in each experiment. The obtained results prove the advantages of the proposed approach.Detecting minimum features and improving accuracy are the advantages. Further, the computational complexity has also been reduced in this approach form the dataset, and also improved accuracy of the applied Naïve Bayes Classifier.

Principal Component Analysis approach is another significant work in this area. This is proposed by Chabathula et.al, [66]. They have taken help of Machine for feature reduction. Effective selection is also one of the objectives of this approach. A different method was considered for performing network analysis for the purpose of reducing the data features. The header fields of incoming packets are processed for analysis in the form of vectors, these serve as import to the PCA algorithm. Two experimental analyses were carried out, one with PCA and another without PCA. Speed, Detection accuracy and the time taken for detection are measured as performance metrics. The detection rates are observed for KNN, Random Forest Tree, SVM, Nearest Neighbor, Adaboost and the J48 tree [67]. They were almost similar for both the cases. These cases are with PCA and without PCA.

This analysis was carried out on Dimension Reduction Technique. Classifier Combination and performance analysis for IDS was considered for such analysis. This approach is proposed by Dhafian et.al, [64] and Chauhanet.al, [45]. Various search methods, Dimension Reduction Techniques, Classifiers and Attribute evaluators are analyzed by authors. Various types of feature selection and feature classification algorithms are developed and applied over Kyoto 2006+, KDD, NSL-KDD, DARPA and CAIDA datasets. The obtained results revealed that the NSL-KDD dataset gives better results once trained against particular classifiers. The improvement in the accuracy observed is declared as from 52% to 96% of PCA.Researchersrecommend that classification with a good accuracy results in a reduction in computational time and effective outputs.

3.1.2.2 Feature Selection

Feature selection is described as a method whereby specific features are selected from a set of features, which have a high discrimination capability between class labels. The main objective behind the feature selection is to reduce the irrelevant features by eliminating them through some techniques. Feature selection is used to maximize performance by reducing the irrelevant features by selecting and ignoring sets at minimum length. While doing so, the attributes with maximum accuracy are kept at the top level.

Ganapathy et.al, [46] proposed aninnovativeIDSusing two separate algorithms. Initially a review was carried out over the existing feature selection and classification approaches and also a survey has been carried out on the intelligent approaches. A new classification approach, named as IREMSVM was developed from the multiclass SVM. This SVM is based on current intelligent agent. Recently extracted features use information gain ratio along with feature selection approach. The aim of this process is the selection of appropriate feature set. Two new approaches, the IREMSVM and IAEMSVM were then processed for testing against the KDD dataset utilizing the complete features and one with features selected. The proposed approach was tested over the DOS, probe and some other attacks. The results obtained declared that the IREMSVM has higher accuracy compared to IAEMSVM or SVM.

Zargariet.al, [39] developed a significant feature selection approach to perform anomaly detection with the aim of applying data mining techniques. This approach used rough set theory to detect the features that are more discriminative in every class. For the discovery of features the amended KDD dataset was considered. Whether there is an improvement with the use of this process is a point of significance. The important thing to be noted is this approach used a corrected KDD dataset instead of KDD dataset. These two datasets are differentiated in terms of number of attacks. In comparison, the corrected KDD dataset has an increased number of attack distribution. The obtained features were tested over the NSL-KDD and declared the higher detection accuracy rates.

Aparicio-Navarro et.al, [50] have come out with a discovery. In this discovery, there are three scenarios with a need for correctly labeled datasets. For an unsupervised IDS, the dataset which was trained must be labeled. The authors developed an approach in which the traffic is labeled automatically. The resulting labeled dataset is a subset of the original unlabeled dataset. There is a possibility for valid information in the remaining data set. The Genetic Algorithm (GA) was used for the feature selection on the new labeled dataset. The implementation of GAis carried out for providing metrics automatically. The intension is the generation of

attack detection results. This is accomplished while reducing false classification risk. The evaluation requires normal conditions. Opposite, it is needed to know the nature of analyzed information. It is of no significance during intrusion detection. The necessity is confined only to the extent of the evaluation of IDS efficiency.

Zhang et.al, [47] perform a review over the current feature selection approaches to formulate an efficient solution in a bid. At first, the Chi-square technique was applied to obtain the most important twenty features in the NSL-KDD dataset. Further, the proposed model applies the Bayesian Network classification to select the features. The authors perform tests on all records in the training set and on the test data NSL-KDD data set with 10-fold cross validity. Bayesian networks are widely accepted as a belief that a model is suitable for working under uncertainty.

Relanet.al, [68] has proposed two techniques. They are "C4.5 Decision tree algorithm and C4.5 Decision tree with Pruning", using feature selection. The second pruning decision tree approach considered only the discrete values attributes of classification. The proposed feature selection approach was tested over both the KDD-dataset and NSL-KDD dataset to train and test. The obtained results declared that the pruning decision tree approach had better results with approximately 98% accuracy.

Amiriet. al [38]developed a simple and effective feature selection technique according to mutual information technique. The authors investigated both linear correlation and mutual information and the proposed method resulted in better accuracy especially for the minority attacks.

Sumiyathaseenikram et.al, [80] proposed a new feature selection approach based on the chi-square and Multi class SVM. However, the main drawback with chi-square based feature selection is that it does not provide much information about the relationship between the features. If the relation between the features of different intrusions is not known, entire dataset needs to be analyzed with the incoming intrusion features by which the processing time increases.

Table.4. Comparative Analysis of various Feature Transformation techniques

Methodology used	**Listed Approaches**	**Advantages**	**Disadvantages**
Feature Reduction	Wahba et.al [63] Tesfahun et.al. [44] Dhafian et.al. [64] Desale et.al. [65] Chabathula et.al. [66] Chauhan et.al. [45]	Reduces the computational overhead due to the smaller feature size.	Decreases the detection accuracy in the case of rare attacks.
Feature Selection	Ganapathy et.al. [46] Zargariet.al. [39] Aparicio-Navarro et.al. [50] Zhang et.al. [47] Relanet.al. [68] Sumiya et.al. [80] Amir et.al [38]	Improved accuracy due to the more discrimination between the selected features.	More Computation times. Less exploration about the feature relations

3.2 Detection Approaches

After obtaining sufficient features, these are processed to classifier to perform anomaly detection. Detection approaches perform the comparison between the testing and the trained features. Based on the obtained results, it decides whether the incoming packet is of intrusion or not. Various types of Detection approaches are developed earlier and brief summary is illustrated in this section.

Broadly used Detection approaches include-

1. Supervised Learning (Classification)
- Decision Tree (DT)
- Naïve Bayes (NB)
- Bayesian Network (BN)
- Logistic Regression (LR)
- Neural Networks (NN)
- Support Vector Machines (SVM)

2. Unsupervised Learning (Clustering)
- K-Means
- CLIQUE

3. Evolutionary Approaches
- Genetic Algorithm
- Particle Swarm Optimization
- Ant Colony Optimization
- Artificial Immune System

3.2.1 Supervised Learning

In supervised learning, the readily available training dataset is pointing out with its target vector. The output vector is obtained after the learning from the available data by taking guidance.

3.2.1.1 Decision Tree Based Approaches

A decision tree resembles the structure of a tree which has leaves and branches. It represents the classifications, which inturn the conjunctions of features those are the result of classifications. An exemplar is labeled (classified) by testing its feature (attribute) values against the nodes of the decision tree. ID3 [51] and C4.5 [68], [4] are the most well-known decision tree based classifiers.

Lydia et.al, [52] proposed a novel IDS based on the "Correlation based Partial Decision Tree Algorithm (CPDT)". The proposed approach extracts the features based on the correlation and used a Partial Decision Tree (PART) for the classification. KDD dataset is used for the performance evaluation and the obtained results shown that the proposed CPDT outperforms the conventional approaches.

In the proposed system by Kajal et.al, [81], the decision tree algorithm is developed based on C4.5 decision tree approach. Two issues are addressed in this approach. Using the information gain ratio, the most relevant features are selected and the separating value is chosen in such a way that the classifier often makes the fairness of the values most often. NSL-KDD dataset was used for the performance evaluation.

The advantages of decision trees are natural learning expression, high classification accuracy, and simple usage. The main disadvantage is that for data including categorical variables with a different number of levels, information gain values are biased in favor of features with more levels. The rule decision for a wide or deeper tree constitutes more complexity. Larger trees frequently have high classification accuracy yet not speculation capabilities.

3.2.1.2 Naïve Bayes

Naïve Bayes is a simple classifier works based on the Bayes theorem. This classifier works based on the assumption of feature independence. i.e., the features are assumed to be independent. Irrespective of the features being they are categorical or continuous, this classifier handles an arbitrary number of independent features. The main advantage with this classifier is its assumption only by which the dimensionality density estimation of the data is reduced.

Saurabh et.al, [40] proposed anIDSbased on the Naive Bayes classifier. The purpose [40] is to identify important reduced input features that are computationally efficient and effective. This approach developed a feature reduction method based on their vital nature to detect the more important reduced input features. Then the Naïve Bayes classifier is applied over the reduced feature set to perform anomaly detection.

V. Hema et.al, [70] proposed a new traffic classification scheme based on Naïve Bayes classifier to discriminate the traffic into intrusion and non-intrusion. Here the network traffic is described through the discretized statistical features and the required information is extracted from these features. Even this approach detects the attacks in the uncertain conditions because the classification methodology is based on the posterior conditional probabilities.

3.2.1.3 Bayesian Network (BN)

Xiaoyan et.al, [69] developed a new IDSbased on the Bayesian Network and PCA. Initially the characteristic values of the attack data of network are evaluated and the main properties are extracted through PCA. Take the main attribute as a new feature set and related principal component contribution rate for improving the traditional characteristic Naive Bayesian classification algorithm. However, the main drawback with PCA is the reduction in the accuracy. To overcome this issue, Tareek M Pattewar et.al, [71] proposed a new IDSbased on Kernel principal component analysis (KPCA) and the Bayesian Network. Working of these techniques are based on neural network. KPCA is an extended version of PCA. After the feature extraction through KPCA [53], the obtained features are processed to Bayesian Network to detect the victim packets.

Ways to detect anomaliesover the attack on the basis of statistical feasibility, which allows for generalization and helps in detecting novel attacks. However, statistical anomaly is not based on a friendly intelligent model and cannot learn from normal and malicious network traffic flow patterns. Alma et.al, [30] developed a network IDS based on the Bayesian network. In this approach, the dataset used for training is a mixed dataset consisting of DARPA dataset and real-world traffic. This IDS model is developed to perform

novel attacks detection. This approach parameterizes the features through the network connections. To test the performance of this approach, the real world dataset and the standard DARPA dataset were utilized. The intrusions which were detected at the first instance are trained to system to increase the effectiveness of system by detecting future intrusions.

The main advantage of Bayesian Network is that the training time is in the linear fashion and it is an online algorithm. Developing a general Bayesian network is an NP Complete Problem, however the simple network of a root node and leaves for attributes from the Naïve Bayes classification structure.

3.2.1.4 Logistic Regression

Logistic Regression is a retrograde model where dependent variables (DV) are categorical. The binary logistics model is used to predict the binary response possibility based on one or more independent (or predictor) variables. The use of LR is carried out in methods of machine learning to classify most data.

Partha et.al, [72]developed an efficient IDSby selecting appropriate features using the LR classifier from NSL-KDD dataset. A new feature selection method called as "Best Feature Set Selection (BFSS)" is proposed based on the genetic algorithm to reduce the learning time and memory space. Then the obtained reduced feature set is subjected to classification through the logistic regression classifier.

Basant et.al, [73]proposed a new anomaly basedIDSusing the LR as a classifier and the linear discriminant analysis (LDA) as feature dimensionality reduction method. NSL-KDD data set is used as a benchmark dataset for the experimental evaluation of this approach and the obtained results revealed that detection rate and accuracy of the both LDA and LR models are too far and even better than the other IDS models.

However, the main disadvantage with LR detection model is the independency between the observations. If the observations are related to one another, the model tends to overweight the significance of those observations. One more issue is the LR model cannot predict the continuous outcomes.

3.2.1.5Neural Networks (NN)

A promising research in Intrusion Detection, for the anomaly identification model, is related to the application of neural network technologies. In order to emulate the operation of the human brain, neural networks have been adopted in the area of inconsistency intrusion detection, primarily due to their adaptability and flexibility for the variations in the environment.

Fatemeh et.al, [74] developed an alert basedanomaly IDSbased on the Neural Networks by which the data is classified into intrusions and non-intrusions in real time and also detects the false positive alerts. Some additional methodologies like preprocessing and filtering are also accomplished in this model to enhance the accuracy.

Vrushali et.al, [82] introduce the Anomaly IDSthat detects various network attacks. This approach aims to detect the network attacks based on the "Back Propagation Artificial Neural Network (BPNN)" algorithm and to protect the system from various attacks. The proposed BPNN algorithm [48] by Jaiganeshet.al, is a supervised neural network. This approach detects the attacks based on their behavior. Initially this approach continuously monitors the behavior and then it decides whether it is an intrusion or not. NSL-KDD dataset was used as a benchmark dataset to test effectiveness and feasibility of the proposed approach. The obtained results revealed that the proper selection of features not only improves the performance and also increases the execution efficiency. The proposed approach also achieved a reduced training time period.

A probabilistic approach is applied by some researchers.Bukhtoyarovet.al, [54] are one of them. They made use of this in the designing of base neural network classifiers termed as"probability-based generator of neural networks structures (PGNS)". The objective is to perform network intrusion detection. This approach was applied over the benchmark KDD dataset with the aim to detect the intrusions belonging to PROBE attacks. It is also used in the detection ofNon-PROBE attacks utilizing the 9 of the 41 attributes. The obtained results declared that the proposed approach with PGNN detected the PROBE attacks effectively.

Jayakumar et.al, [75] proposed a novel IDS based on the Neural Networks(NN) and a feature selection approach. This approach considered only the relevant features instead of considering all the features using the supervised learning neural network to detect the specific attack. Since only the relevant features are considered for intrusion detection, the time taken for processing is reduced more effectively through this approach. Genetic Algorithm and Information Gain Algorithm are used in this model for the feature selection. To train the feature set to the system, "Multi-Layer Perceptron (MLP) supervised NN" is used. Compared to the conventional approaches which use all the features with genetic algorithm and MLP-NN approach, this approach observed to obtain an increased detection rate. This approach achieved an improved detection rate for various network attacks especially for User to Root (U2R) and Remote to Local (R2L) and DOS attacks.

In [83], a proposed system has been developed by Ibraim et.al, that achieves classification technique by using hybrid soft computing technique which is Multi-Layer-Perceptron (MLP) with Particle Swarm

Optimization (PSO). The main aim of PSO is to increase the learning capacity of MLP-NN by setting up the linkage weights in an attempt to improve the classification accuracy of MLP-NN. Further the simulation results described that the classification accuracy of this approach is high when compared to other related approaches.

Though the NN based IDS archives better results in the detection of various attacks, it has significant limitations. The main drawback relates to the training requirements of the neural network. Because the ability of the artificial neural network to identify indications of an intrusion is completely dependent on the accurate training of the system, the training data and the training methods that are used are critical. The training routine requires a very large amount of data to ensure that the results are statistically accurate. The training of a neural network for anomaly detection purpose require thousands of individual attack sequences, and this quantity of sensitive information is difficult to obtain.

3.2.1.6 Support Vector Machine (SVM)

The SVM classifier works in light of finding an isolating hyper plane in the feature space. This is accomplished between two classes. This approach works such that the separation between the hyper plane and the nearest data points of each class is augmented. The approach depends on a limited arrangement chance [37] as opposed to on optimal classification.

Manjiri et.al, [76] developed a new IDSbased on the SVM to classify the attacks in form the raw intrusion datasets for standard personal computers. SVM is methods which perform the data classification based on some vectors which are close to the predefined hyperplane. Here the standard KDD dataset is used for performance testing.

Senthilnayaki et.al, [77]developed a novel IDSbased on the SVM and GA. Here the SVM is used as a classifier and the genetic algorithm is for feature selection. The obtained results reveal that the proposed new feature selection approach using genetic algorithm based on the SVM classification gives better results. This is due to the fact that the proposed approach of feature selection improves the classifier performance in the attack detection by extracting the most useful attributes and also reduces the false alarm rate effectively.

Kabiret.al, [55] proposed a new intrusion detection framework based on "sampling with Least Square Support Vector Machine (LS-SVM)". The entire decision making is carried out in two phases. In the first phase, the entire dataset is divided into some predefined arbitrary subgroups. The proposed algorithm selects the sample of these subgroups, such as samples reflect the whole dataset. An optimum allocation method has been developed based on the variability in the observations within the existing subgroups. The proposed LS-SVM is applied over the obtained samples in the second phase to perform intrusion detection. All binary classes are tested and the proposed approach achieves realistic performance in terms of efficiency and accuracy.

Adriana et.al, [56] proposed an IDS model based on Information Gain for feature selection combined with the SVM classifier. The parameters of SVM are selected by "Swarm Intelligence Algorithms (Particle Swarm Optimization or Artificial Bee Colony)". NSL-KDD data set was used for the performance evaluation and the obtained results reveal that the proposed approach has less false alarm rates and high detection rates than the conventional SVMs.

One possibleadvantage of this approach is to maximize classifier generalization and minimize the bias in the KDD dataset. However, the main disadvantage with SVM is the performance of SVM classifier in the case of known attacks is not effective whereas for real time data, it outperforms the all the existing techniques.

3.2.2 Unsupervised learning (Clustering)

Unsupervised learning systems learn from their environment. Since there is no availability of target vector, systems learn from training data.

Clustering [2] is a data grouping technique in which the data with similar characteristics is arranged into a single group. Particularly the approaches working over the unlabeled data prefers the clustering techniques. This is the method of finding an unsupervised pattern where the data is grouped together based on the similarity measurement. The main benefit of clustering for detection of intrusion is that it can learn from audit data without providing a clear description of the various attack classes to the system administrator. Earlier various clustering models are proposed to cluster the input data. In connectivity based models (e.g., hierarchical clustering), the clustering is based on the distance measurement between them. In centroid models (e.g., k-means), every cluster is designated with its mean vector. In Distribution, Based Models (e.g., Expectation Maximization algorithm), the grouping is taken based on the assumption of statistical distribution acquisition. Density models group the data points as dense and connected regions (e.g., Density-Based Spatial Clustering of Applications with Noise [DBSCAN]). Finally, graph basedmodels (e.g., clique) define every cluster as a set of connected data points where each data point has an edge to at least one other data point in the set.

3.2.2.1 K-Means

Naila et.al, [78] propose a clustering-based anomaly detection technique using a genetic algorithm named "Genetic Clustering for Anomaly-based Detection (GC-AD)". To formulate the clusters, GC-AD used the dissimilarity measure. Then it applies a genetic process where every chromosome represents the centroids of k clusters. The standard KDD dataset was used for testing the proposed approach and the obtained accuracy results are compared with the k-means clustering approach.

Mohsen et.al, [58] proposed a new intrusion detection method using "Min Max K-means clustering algorithm", which overcomes the shortage of sensitivity to initial centers in K-means algorithm, and increases the quality of clustering. The experiments observations on the NSL-KDD data set designate that the proposed clustering approach is more efficient than the most conventional K-means clustering approach.

Sita et.al, [51] examined several Similarity/Dissimilarity methods for Intrusion Detection issue. An offline incompatibility based IDSwas implemented using the agglomerative and partial based clustering algorithm. Two cluster labeling algorithms were employed, Similarity Normal Clustering labeling algorithms and class representative objects used to label the groups using objects. KDD dataset was used for evaluation.

Though the clustering techniques archive an optimal performance, they have significant limitations. The main disadvantage with K-means is its dependence on initial centroids, dependence on number of clusters and degeneracy.

3.2.2.2 CLIQUE

There are a few proposed grouping calculations to identify inconsistency on IDS, yet the greater part of them discover bunches in the most noteworthy measurement of information. Club Partitioning (CP) [57] is one of the grouping calculations proposed by Nasta et.al, [57] that discovers bunches from the subspace of information. In view of computational time, flawlessness and false caution rate, the framework is tried to dissect the execution. The proposed CP approach accomplished great execution in the perspective of fulfillment (94.59%) and false caution rate (2.54%). This approach also attained an optimal performance in the view of computational time based on the tuple, however not achieved an efficient performance in the view of quantity.

Table.5. Comparative Analysis of various Classification Techniques

Method	**Listed Approaches**	**Advantages**	**Disadvantages**
Decision Tree	Lydia et.al. [52] Kajal et.al. [81] Sahu et.al. [67]	Intuitive knowledge expression. High classification accuracy. Simple implementation.	For a tree with deeper level, the rule based decision constitutes more complexity.
Naïve Bayes	Saurabh et.al. [40] V. Hema et.al. [69]	Easy to implement. Reduces the dimensions of dataset effectively.	Assumes that the variables are independent.
Bayesian Network	Xiaoyan et.al. [70] Pattewar et.al. [71] Alma et.al. [30]	Online algorithm and its training is completed in linear time.	Bayesian network construction is an NP complete problem.
Logistic Regression	Partha et.al. [72] Basant et.al. [73]	Capable of handling non-linear data. More robust.	Independency between the observations. Cannot predict the continuous outcomes.
Neural Networks	Fatemeh et.al. [74] Vrushali et.al. [82] Jaiganesh et.al. [48] Bukhtoyarov et. al. [54] Jayakumar et.al. [75] Ibraim et.al. [83]	High Accuracy. Noise tolerance. Independence from prior assumptions. Ease of maintenance. Flexible Implementation.	Training requires bulk data to ensure the accuracy. No method to determine optimal neurons.
Support Vector Machines	Manjiri et.al. [76] Senthilnayaki et.al. [77] Kabir.et.al. [55] Adriana et.al. [56]	Maximizes Classifier generalization. Minimizes the bias in the KDD data set.	Though works effectively for real time data, not effective for known attacks.
Clustering	Nasta et.al. [57] Naila et.al. [78] Mohsen et.al. [58] Sita et.al. [51]	Reduced Data set size. Natural grouping even for unlabeled data. Intra cluster similarity is high.	Dependence on initial centroids. Determination of number of clusters is complex.

3.2.4 Evolutionary Approaches

The term evolutionary computation encompasses Genetic Algorithms (GA) [3], [59], [41],Evolution Strategies [14], Particle Swarm Optimization [5],Ant Colony Optimization [6], and Artificial Immune Systems[1].

3.2.4.1 Genetic Algorithm

Salah et.al, [60] developed an intelligent IDSbased on the Genetic Algorithm(GA) with an enhanced initial population and selection operator, to perform various attacks effectively. Here the main use of GA is to optimize the search of the attack scenarios in the audit files. It tends to find the potential attacks in the audit files. The NSL-KDD benchmark dataset was used in the testing phase to measure the detection rate of proposed approach. The obtained results reveal that the combination of Genetic algorithm with the IDS improved the detection rate and also reduces the false alarm rate. A new intrusion detection approach proposed by Dheeraj pal et.al, [61] also reveal that the combination of Genetic algorithm with IDSbased on the information gain attains a reduced feature set with an improved performance. Thus, the proposed approach attained a reduced computational complexity. Apart from this, a soft computing approach was built in the rule creation, which makes governance more efficient from the hard computing approach used in the current genetic algorithm.

3.2.4.2Particle Swarm Optimization

Ibrahim et.al, [49] proposed an IDSbased on a parallel particle swarm optimization utilizing the MapReduce methodology. Using particle swarm optimization for clustering work, particle swarm optimization is a very effective method because the particle swarm optimization avoids sensitivity problems of initial cluster centroids as well as premature convergence. The proposed IDSset large data on the object hardware. Experimental results on actual intrusion data set shows that the proposed IDSis very well scalable with increasing sizes of data.

Yang et.al, [84] proposed an "Improved Particle Swarm Optimization algorithm ICPSO", which uses chaos operator periodicity, randomness, sensitivity to initial conditions and other characteristics. The proposed ICPSO is used to make the chaos into the inertia weight factor parameters. The chaos is applied to obtain the RBF kernel factor optimization and also the penalty parameter. The proposed ICPSO also aims to achieve an increased precision and convergence speed of the PSO. Simulation results reveal that the proposed ICPSO-SVM outperforms the conventional GA-SVM and PSO-SVM.

3.2.4.3Ant Colony Optimization

The purpose of Mehendi et.al, [85] is to identify important features in building an IDSsuch that they are computationally efficient and effective. In order to improve the performance of the IDS, [85] an IDSoffers that its features have been better chosen with the use of ant colony optimization [86]. Due to the use of fixed simplified facility for classification, the proposed method is easily applied and has less computational complexity. The extensive experimental results on the KDD and NSL-KDD intrusion detection benchmark data sets demonstrate that the efficiency of proposed method.

3.2.4.4 Artificial Immune System

Eman et.al, [79] utilized artificial immune system network based intrusion detection. In the proposed structure GureKddcup database set is utilized for intrusion detection and utilized R-chunk algorithm [86] of artificial immune framework strategy, it is utilized for anomaly detection. An optimized feature selection of rough set hypothesis utilized for improving tedious.

Obinna et.al, [87] presents a system for a "Distributed Network Intrusion Detection System (DNIDS)" based on the artificial immune system strategy. In this approach, the adaptive immune framework is proposed to classify network traffic in ordinary and suspicious profiles, separately, through strategies for machine learning. The experimental proposed approach delivers NIDs to all connected network segments, allowing NIDs to be able to identify different potential threats in each section and to share threatened vectors between distributed NIDS.

IV. CONCLUSION

In this report, a brief literature survey is carried out the over the approaches developed earlier for Intrusion Detection. A comprehensive survey of Data Preprocessing techniques such as Basic Features based approaches, TSF based Approaches, TSMF based approaches are explored. Further the obtained features are subjected to Feature Transformations such as Feature Extraction/Reduction and Feature Selection. Various approaches based on these feature transformations are also explored in this report. Further the survey is carried out over the IDS mainly focusing on the supervised learning classifiers such as Decision tree, Naïve Bayes classifier, Bayesian network, Logistic Regression, Neural Networks and Support vector machine and also over

unsupervised learning approaches like K-means and CLIQUE. Some of Evolutionary Approaches like Genetic Algorithm based IDS, PSO based IDS, ACO based IDS and Artificial Immune System are also illustrated in this report. The report contains the possible Pros and Cons which are explored at each and every stage.

One important observation of this survey is that most of the authors focused towards the detection approaches such as supervised and Unsupervised Learning. Most of the them are developed based onANN, SVM, Fuzzy, K-means clustering etc. Though this approach obtained better detection accuracy, feature selection also needs to be considered after which the detection takes place. Since the standard datasets such KDD and NSL-KDD have a fixed set of features, most of the IDS models tried to apply detection algorithms over the entire dataset by which the computational overhead increases followed by an increased computational time. Hence there is a need to construct efficient feature selection/extraction techniques for outlier detection by classify them through a learning cluster based statistical mapping approaches for computing IDS in Machine Learning approaches, by which the redundant data reduces with less loss of information.

REFERENCES

[1]. J. Farmer, N. Packard, and A. Perelson,The immune system, adaptation and machine learning,Phys. D: Nonlinear Phenomenon, 2, 1986, pp. 187–204.
[2]. K. Jain and R. C. Dubes, Algorithms for Clustering Data,EnglewoodCliffs, NJ, USA: Prentice-Hall, 1988.
[3]. D. E. Goldberg and J. H. Holland, Genetic algorithms and machinelearning,Machine Learning, 3(2),1988, pp. 95–99.
[4]. R. Quinlan, C4.5: Programs for Machine Learning, San Mateo, CA, USA: Morgan Kaufmann, 1993.
[5]. J. Kennedy and R. Eberhart, Particle swarm optimization, IEEE International Conference on Neural Networks, IV, 1995, pp. 1942–1948.
[6]. M.Dorigo and L. M. Gambardella, Ant colony system: A cooperative learning approach to the traveling salesman problem,IEEE Transactions on Evolutionary computing, 1(1), 1997, pp. 53–66.
[7]. S. Axelsson, Research in intrusion-detection systems, a survey, Department of Computer Engineering, Chalmers University of Technology, Goteborg, Sweden, Technical Report, 1998, 98-17.
[8]. Dickerson J, Dickerson J, Fuzzy, Network profiling for intrusion detection, 19th International Conference of the North American Fuzzy Information Processing Society (NAFIPS), 2000, pp. 301–306.
[9]. Sans Penetration testing, Host-vs. Network based Intrusion detection systems, 2001.
[10]. Mahoney M, Chan P, PHAD: Packet header anomaly detection for identifying hostile network traffic, Florida Institute of Technology technical report CS-2001-04.
[11]. Barbara D, Wu N, and Jajodia S, Detecting novel network intrusions using Bayes estimators, First SIAM Conferenceon Data Mining, 2001.
[12]. Northcutt, S. & Novak, J, Network intrusion detection, 2002, Sam's Publishing.
[13]. Staniford S, Hoagland J, Practical automated detection of stealthy portscans, Journal of Computer Security, 10(1), 2002, pp.105–36.
[14]. H. G. Beyer and H. P. Schwefel, Evolution strategies: A comprehensiveintroduction,J. Nat. Computing, 1(1), 2002, pp. 3–52.
[15]. Ramadas M, Tjaden B, Detecting anomalous network traffic with self-organizing maps, Lecture Notes in Computer Science 2003, pp.36–54.
[16]. Estevez-Tapiador JM, Garcia-Teodoro P, Stochastic protocol modelling for anomaly based network intrusion detection, International Workshop on Information Assurance, IWIAS 2003, pp. 3–12.
[17]. Lazarevic A, Ertoz L, Kumar V, Ozgur A, Srivastava J, A comparative study of anomaly detection schemes in networkintrusion detection,3rdSIAM International Conference on Data Mining, 2003, pp. 25–36.
[18]. Ertoz L, Lazarevic A, Kumar V, Srivastava J, Minds-Minnesota intrusion detection system,Next Generation Data Mining, 2004.
[19]. Hamdi, M. &Boudriga, N, Computer and network security risk management: Theory, challenges, and, International journal of communication systems, 18 (8), 2005, pp.763-793.
[20]. Molsa, J, Mitigating denial of service attacks, a tutorial, Journal of computer security, 13 (6), 2005, pp.807-837.
[21]. Lakhina A, Crovella M, Diot C, Mining anomalies using traffic feature distributions, International conference on Applications, technologies, architectures, and protocols for computer communications, ACM, 2005, pp.228-235.
[22]. Early J, Brodley C, Behavioral features for network anomaly detection, Machine Learning and Data Mining for Computer Security 2006, pp.107–24.
[23]. Santos, O, End-to-end Network Security, Defense-in-depth. Pearson Education, 2007.
[24]. OnutI, Ghorbani A, A feature classification scheme for network intrusion detection, International Journal of Network Security, 5(1), 2007, pp.1–15.
[25]. Yamada A, Miyake Y, Perrig A, Intrusion detection for encrypted web accesses,21st International conference in Advanced Information Networking and Applications Workshops, AINAW'07, 2007, pp. 569–576.
[26]. Yang J, Huang SHS, Mining TCP/IP packets to detect stepping-stone intrusion, Computers & security, 26(7-8), 2007, pp.479–484.

[27]. Pokrajac D, Lazarevic A, Latecki L, Incremental local outlier detection for data streams, IEEE Symposium on Computational Intelligence and Data Mining (CIDM), 2007.

[28]. Patcha A, Park JM, Network anomaly detection with incomplete audit data, Computer Networks, 51(13), 2007a, pp.3935–55.

[29]. Guennoun M, El-Khatib K, Selecting the best set of features for efficient intrusion detection in 802.11networks, International conference on Information and Communication Technologies: From Theory to Applications (ICTTA),2008, p. 1–4.

[30]. Alma, Network Intrusion Detection Based on Bayesian Networks, 20th International Conference on Software Engineering & Knowledge Engineering (SEKE'2008), San Francisco, CA, USA, 2008.

[31]. P.García-Teodoro, Anomaly-based network intrusion detection: Techniques, systems and challenges, Computers & Security,28(2), 2009, pp.18-28.

[32]. Zhao J, Huang H, Tian S, Zhao X, Applications of HMM in protocol anomaly detection, International Joint Conference on Computational Sciences and Optimization(CSO-2009), 2009, pp. 347–349.

[33]. Lu W, Ghorbani A, Network anomaly detection based on wavelet analysis, EURASIP Journal on Advances in Signal Processing,2009, pp.4–10.

[34]. Evans, D, The Internet of Things. How the Next Evolution of the Internet is Changing Everything, Whitepaper, Cisco Internet Business Solutions Group (IBSG), 2010.

[35]. Wu, Q., Shiva, S., Roy, S., Ellis, C., &Datla, V, On modeling and simulation of game theory based defense mechanisms against DoS and DDOS attacks, In Proceedings of the 2010 spring simulation multi-conference,2010, pp. 159.

[36]. Muraleedharan N, Parmar A, Kumar M, A flow based anomaly detection system using chi-square technique,IEEE 2nd International on Computing Conference (IACC), 2010, pp. 285–9.

[37]. V. Vapnik, The Nature of Statistical Learning Theory, New York, USA, Springer, 2010.

[38]. Amiri, Fatemeh, Mahdi, Mohammad, Mutual information based feature selection for intrusion detection systems. J. Network Computers, 34, 2011, pp.1184–1199.

[39]. S. Zargari and D. Voorhris, Feature Selection in the Corrected KDD-dataset,3rd International Conference on Emerging Intelligent Data and Web Technologies, 2012, pp. 174-180.

[40]. Dr. Saurabh Mukherjee, Intrusion Detection using Naive Bayes Classifier with Feature Reduction, Elsevier, Procedia Technology, (4), 2012, pp. 119 – 128.

[41]. S Ganapathy, A Novel Weighted Fuzzy C-means Clustering Based on Immune Genetic Algorithm for Intrusion Detection, International Conference on Modeling Optimization and Computing, 2012.

[42]. Philip Chan, Signature Based Intrusion Detection Systems, Springer 2013.

[43]. F. Zhang and D. Wang, An Effective Feature Selection Approach for Network Intrusion Detection,IEEE 8th International ConferenceinNetworking, Architecture and Storage, 2013, pp. 307-311.

[44]. A. Tesfahun and D.L. Bhaskari, Intrusion Detection using Random Forests Classifier with SMOTE and Feature Reduction,International Conference on Cloud & Ubiquitous Computing & Emerging Technologies, 2013, pp. 127-132.

[45]. H. Chauhan, V. Kumar, S. Pundir and E.S. Pilli, A Comparative Study of Classification Techniques for Intrusion Detection, IEEE International Symposium on Computational and Business Intelligence,2013, pp. 40-43.

[46]. Ganapathy. S, Intelligent Feature Selection and Classification Techniques for Intrusion Detection in Networks: A survey,EURASIP Journal on Wireless Communications and Networking,1(271), 2013, pp. 1-16.

[47]. F. Zhang and D. Wang, An Effective Feature Selection Approach for Network Intrusion Detection, IEEE Eighth International Conference on Networking, Architecture and Storage,2013, pp. 307-311.

[48]. V. Jaiganesh, P. Sumathi, and S. Mangayarkarasi,An analysis of intrusion detection system using back propagation neural network, International Conference on Information Communication and Embedded Systems (ICICES), 2013, pp. 232–236.

[49]. Ibrahim, MapReduce intrusion detection system based on a particle swarm optimization clustering algorithm, IEEE Congress on Evolutionary Computation (CEC), 2013.

[50]. F. Aparicio-Navarro, K.Gand D.J. Parish, Automatic Dataset labeling and Feature Selection for Intrusion Detection Systems,IEEE Military Communications Conference, 2014, pp. 46 - 51.

[51]. Sita Rama Murthy, Exploring the Similarity/Dissimilarity measures for unsupervised IDS, International Conference on Data Mining and Advanced Computing (SAPIENCE), 2014.

[52]. F.Lydia, Efficient host based intrusion detection system using Partial Decision Tree and Correlation feature selection algorithm, International Conference on Recent Trends in Information Technology (ICRTIT), 2014.

[53]. Kuang, Fangjun, A novel hybrid KPCA and SVM with GA model for intrusion detection, Applied Soft Computing, 2014.

[54]. V. Bukhtoyarov, V. Zhukov, Ensemble-distributed approach in classification problem solution for intrusion detection systems, Intelligent Data Engineering and Automated Learning-IDEAL 2014, Springer, 2014, pp. 255–265.

[55]. M Kabir, A statistical framework for intrusion detection system, 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2014.

[56]. Adriana, Intrusions detection based on Support Vector Machine optimized with swarm intelligence, IEEE 9th International Symposium on Applied Computational Intelligence and Informatics (SACI), 2014.

[57]. Nasta, Anomaly Detection on Intrusion Detection System Using CLIQUE Partitioning, 2nd International Conference on Information and Communication Technology (ICOICT), 2014.

[58]. Mohsen, Intrusion detection based on MinMax K-means clustering, 7th International Symposium on Telecommunications (IST), 2014.

[59]. FangjunKaung, A novel hybrid KPCA and SVM with GA model for intrusion detection, Applied Soft Computing, 18, 2014, pp. 178–184.

[60]. Salah, Intrusion detection system using genetic algorithm, Science and Information Conference (SAI), 2014.

[61]. Dheeraj Pal, Improved Genetic Algorithm for Intrusion Detection System, International Conference on Computational Intelligence and Communication Networks (CICN), 2014.

[62]. Bhavesh Kasliwal, Shraey Bhatia, Shubham Saini, A Hybrid Anomaly Detection Model using G-LDA,IEEE 2014.

[63]. Y. Wahba, E. Elsalamouny and G. Eltaweel, Improving the Performance of Multi-class Intrusion Detection Systems using Feature Reduction,International Journal of Computer Science Issues,12(3), 2015, pp. 355-368.

[64]. B. Dhafian, I. Ahmad and A. AL-Ghamid, An Overview of the current Classification Techniques in Intrusion Detection, in International Conference Security and Management,2015, pp. 82-88.

[65]. K.S. Desale and R. Ade, Genetic Algorithm based Feature Selection Approach for Effective Intrusion Detection System, 3rd International Conference on Computer Communication and Informatics, 2015, pp. 1-6.

[66]. K.J. Chabathula, C.D. Jaidhar, M.A. Ajay Kumara, Comparative Study of Principal Component Analysis based Intrusion Detection approach using Machine Learning Algorithms,3rd International Conference on Signal Processing Communication and Networking,2015, pp. 1-6.

[67]. Sahu, Network intrusion detection system using J48 Decision Tree,International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2015.

[68]. N.G. Relan and D.R. Patil, Implementation of Network Intrusion Detection System using Variant of Decision Tree Algorithm,IEEEInternational Conference Nascent Technologies in the Engineering Field,2015, pp. 1-5.

[69]. Xiaoyan, A Naive Bayesian Network Intrusion Detection Algorithm Based on Principal Component Analysis, 7th International Conference onInformation Technology in Medicine and Education (ITME), 2015.

[70]. V. Hema, DoS Attack Detection Based on Naive Bayes Classifier, Middle-East Journal of Scientific Research on Sensing, Signal Processing and Security, 2015, pp.398-405.

[71]. Tareek M Pattewar, Neural network based intrusion detection using Bayesian with PCA and KPCA feature extraction, IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS), 2015.

[72]. ParthaGhosh, Proposed GA-BFSS and logistic regression based intrusion detection system, 3rd International Conference on Computer, Communication, Control and Information Technology (C3IT), 2015.

[73]. BasantSubba, Intrusion Detection Systems using Linear Discriminant Analysis and Logistic Regression, Annual IEEE India Conference (INDICON), 2015.

[74]. FatemehCharlank, Evaluating Artificial Neural Network in Intrusion Detection System Alert Management System, Journal of Scientific Research and Development,2 (5), 2015, pp. 316-319.

[75]. Kaliappan Jaya kumar, Intrusion Detection using Artificial Neural Networks with Best Set of Features, The International Arab Journal of Information Technology, 12(6A), 2015.

[76]. Manjiri, Classification of Attacks Using Support Vector Machine (SVM) on KDDCUP'99 IDS Database,5th International Conference on Communication Systems and Network Technologies(CSNT), 2015.

[77]. B.Senthilnayaki, Intrusion detection using optimal genetic feature selection and SVM based classifier, 3rd International Conference on Signal Processing, Communication and Networking (ICSCN), 2015.

[78]. Naila, A genetic clustering technique for Anomaly-based Intrusion Detection Systems, 16th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015.

[79]. Eman, Artificial immune system based intrusion detection, IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS), 2015.

[80]. SumiyaThaseen Ikram, Intrusion Detection Model using fusion of chi-square feature selection and multiclass SVM, Journal of King Saud University-Computer and Information Sciences, 2016.

[81]. KajalRai, Decision Tree Based Algorithm for Intrusion Detection, Int. J. Advanced Networking and Applications,07(04), 2016, pp.2828-2834, 2016.

[82]. Vrushali D. Mane, Anomaly based IDS using BackPropagation Neural Network, International Journal of Computer Applications, 136(10), 2016, pp. 0975 – 8887.

[83]. Ibraim M. Ahmed, Enhancement of Network Attack Classification using Particle Swarm Optimization and Multi-Layer-Perceptron, International Journal of Computer Applications,137(12), 2016, 0975 – 8887.

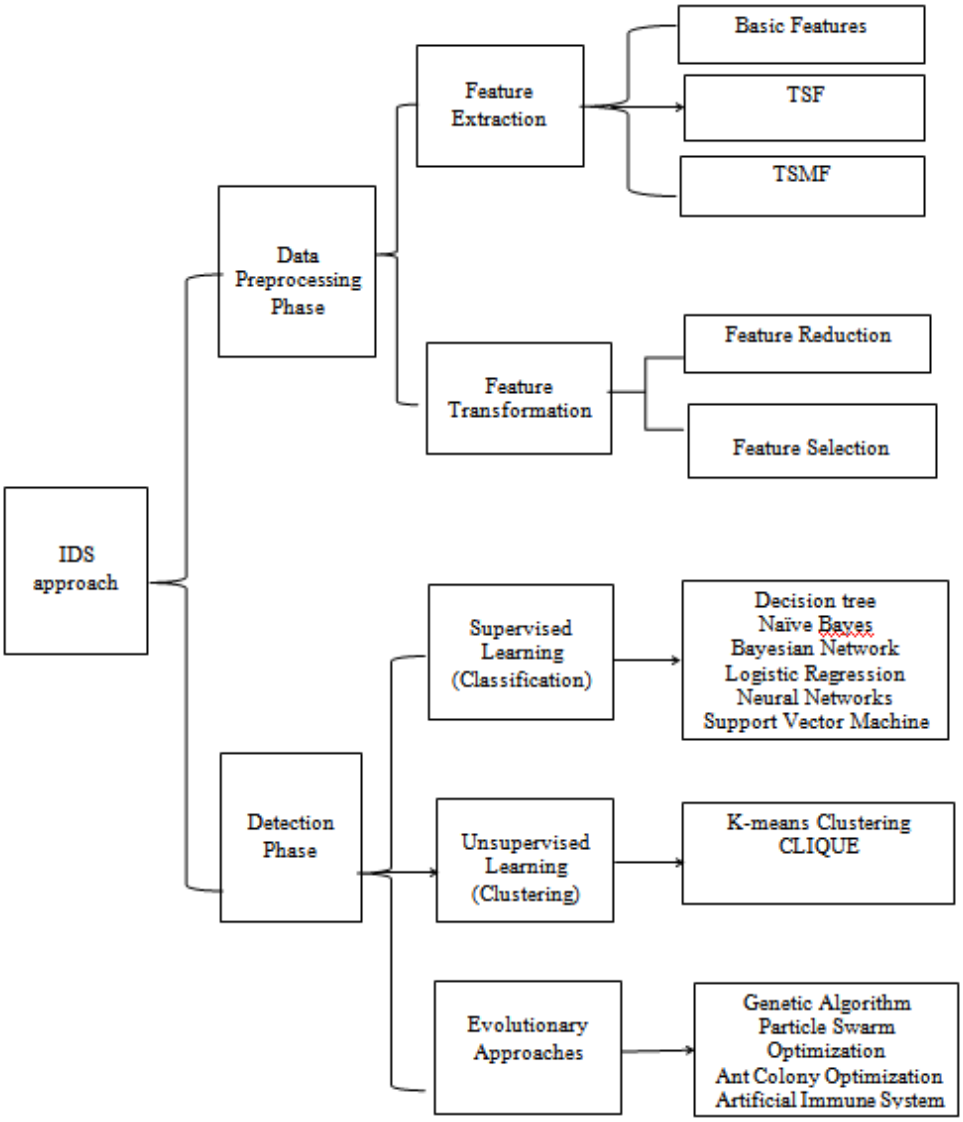
[84]. Yang,An Optimization Method for Parameters of SVM in Network Intrusion Detection System, International Conference on Distributed Computing in Sensor Systems (DCOSS), 2016.

[85]. Mehdi HosseinzadehAghdam, Feature Selection for Intrusion Detection System Using Ant Colony Optimization, International Journal of Network Security, 18(3), 2016, pp.420-432.

[86]. Ravi KiranVarma, Feature Selection using relative fuzzy entropy and ant colony optimization applied to real time intrusion detection system, International conference on Computational Modeling and Security (CMS 2016), procedia computer science, 85, 2016, pp. 503-510.

[87]. Obinna,Ihab Darwish, Tarek Saadawi,Distributed Network Intrusion Detection Systems: An Artificial Immune System Approach, IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2016.

Appendix: Tree representation of the literature survey



B. Narendra Kumar[1]" A Comparative Survey on the Influence of Machine Learning Techniques on Intrusion Detection System (IDS)."IOSR Journal of Engineering (IOSRJEN), vol. 08, no. 8, 2018, pp. 25-42.