Taxonomy of Anomaly Based Intrusion Detection System: A Review

Manasi Gyanchandani*, J.L.Rana, R.N.Yadav*****

* IT Dept. MANIT, Bhopal.
** Ex-HOD, CSE/IT Dept. MANIT, Bhopal.
*** ECE Dept. MANIT. Bhopal.

Abstract- Intrusion detection systems aim at detecting attacks against information systems in general. It is difficult to provide secure information systems and maintain them in a secure state for their entire lifetime. Maintenance of such information system is technically difficult as well as economically costly. With the advent of new vulnerabilities to information system new techniques for detecting these vulnerabilities have been implemented. In this paper we introduce a taxonomy of anomaly based intrusion detection systems that classifies all possible techniques. It defines families of anomaly based intrusion detection systems according to their properties along with their advantages and disadvantages. This paper also includes various examples from the past and current projects. We hope that this survey will provide a better understanding of the different directions in which research has been done on this topic.

Index Terms- Novel attacks, anomalies, false alarms, detection rate, accuracy.

I. INTRODUCTION

Internet has become a part of daily life and an essential tool today. Security of using Internet is one major research problem for business and personal networks. Network based attacks are increasing frequently resulting a huge financial loss to the organizations and causing the network to be paralyzed for several hours. Security threats to the computer systems have raised the importance of intrusion detection systems. Intrusion Detection (ID) acts as a powerful weapon to protect these networks from intrusive activities. In the year 1980 J. P. Anderson introduced the concept that certain types of threats to the security of computer systems could be identified through a review of information contained in the system's audit trail [1]. Denning and Neumann [2] in the year 1985 has recommended the need for effective intrusion detection mechanisms as part of a security mechanism for computer systems. In the year 1987 Dorothy E. Denning introduced the concept of ID, and it has become a landmark in the research area [3]. The model which she proposed has formed the fundamental core of most intrusion detection methodologies in use today.

Intrusion detection is the process of intelligently monitoring the events occurring in a computer system or network and analyzing them for signs of violations of the security policy, Parker [4] has defined six security issues to be considered while designing an IDS: Confidentiality, Integrity, Availability, Utility, Authenticity, and Possession of a computer or network. Intrusions are caused by attackers accessing the systems from the Internet, authorized users of the systems who attempt to gain additional privileges for which they are not authorized, and authorized users who misuse the privileges given them. Intrusion Detection Systems (IDSs) are software or hardware products that automate this monitoring and analysis process.

There are two main types of Intrusion Detection System (IDS): Signature Based IDS (SBIDS) and Anomaly Based IDS (ABIDS). [5]

In SBIDS, also known as misuse detection, signatures of known attacks are stored and the events are matched against the stored signatures. It will signal an intrusion if a match is found. The main drawback with this method is that it cannot detect new attacks whose signatures are unknown. This means that an IDS using misuse detection will only detect known attacks or attacks that are similar enough to a known attack to match its signature.

ABIDS has attracted many academic researchers due to its potential for addressing novel attacks. Novelty detection is the identification of new or unknown data that a machine learning system is not aware of during training [6]. ABIDS have two major advantages over signature based intrusion detection systems. The first advantage is the ability to detect unknown attacks as well as ''zero day'' attacks. This is because of the ability of anomaly detection systems to model the normal operation of a system/network and detect deviations from them. A second advantage is that the aforementioned profiles of normal activity are customized for every system, application and/or network, and therefore making it very difficult for an attacker to know with certainty what activities it can carry out without getting detected [5].

II. TECHNIQUES OF ABIDS

In this subsection we review different techniques of Anomaly based IDS. The most important are Statistical anomaly detection, Data-mining based detection, Knowledge based detection, and Machine learning based detection. The complete taxonomy of ABIDS is shown in the Figure 1.

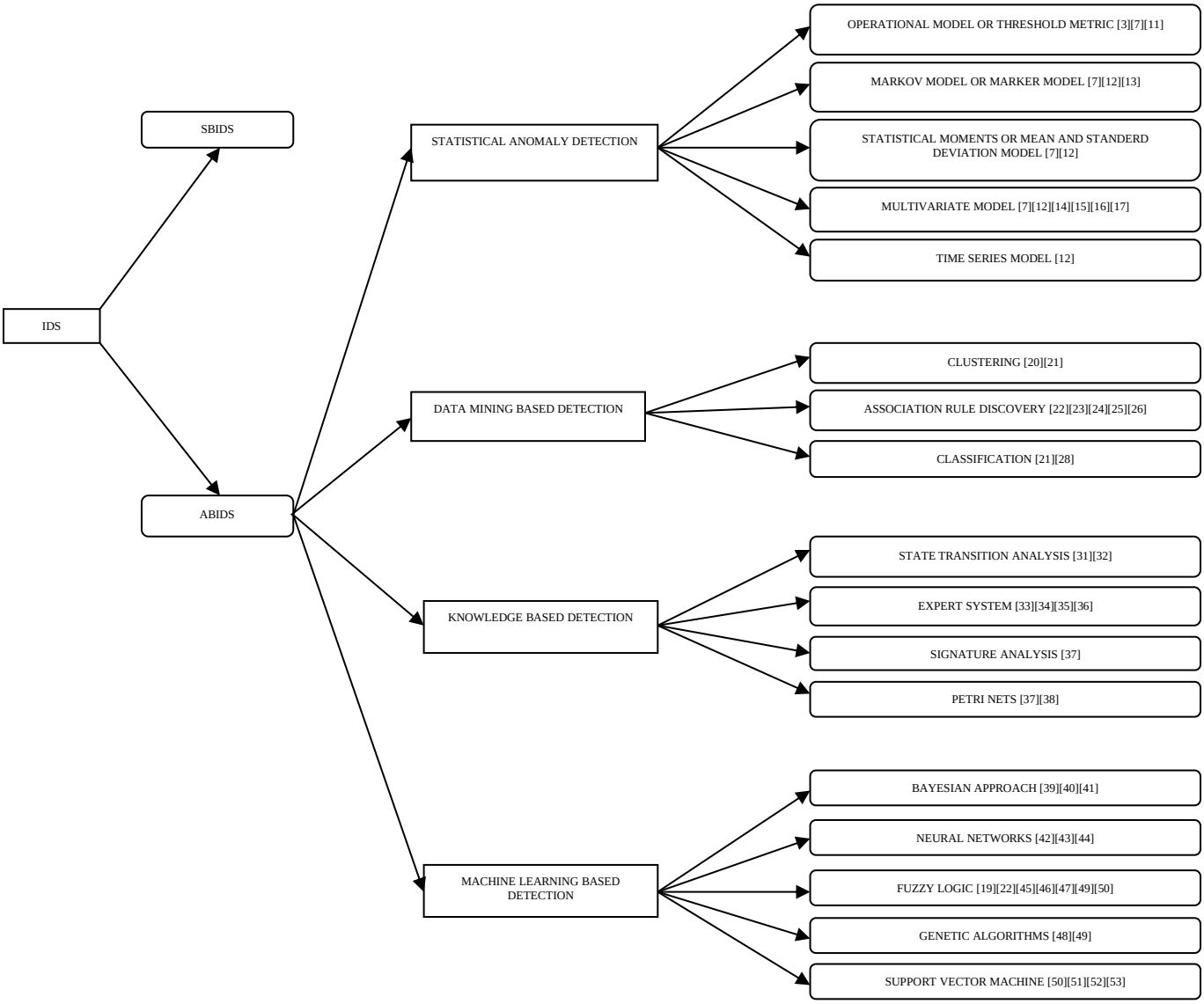

Figure 1: Taxonomy of Anomaly based Intrusion Detection System.

III. STATISTICAL ANOMALY BASED INTRUSION DETECTION SYSTEM (SABIDS)

Statistical modeling is among the earliest methods used for detecting intrusions in electronic information systems. Statistical based anomaly detection techniques use statistical properties and statistical tests to determine whether "Observed behavior" deviate significantly from the "expected behavior" [7]. Statistical based anomaly detection techniques use statistical properties (e.g., mean and variance) of normal activities to build a statistical based normal profile and employ statistical tests to determine whether observed activities deviate significantly from the normal profile. The IDS goes on assigning a score to an anomalous activity. As soon as this score becomes greater certain threshold, it will generate an alarm. SABIDS is a two-step process: first it establishes behavior profiles for the normal activities and current activities. Then these profiles are matched based on various techniques to detect any kind of deviation from the normal behavior. SABIDS can further be classified into following categories: [3] [8]

a. Operational Model or Threshold Metric
b. Markov Process Model or Marker Model
c. Statistical Moments or Mean and Standard Deviation Model
d. Multivariate Model

e. Time Series Model

Advantages of SABIDS [8] [9] [10]

1. SABIDS do not require prior knowledge of security flaws and/or the attacks themselves. As a result these systems can detect "zero days" or the extremely new attack.
2. Statistical approaches can provide accurate notification of malicious activities that typically occur over extended periods of time and are good indicators of impending denial-of-service attacks.
3. These systems are potentially easier to maintain, as there is no need to update signatures and the system doesn't rely on specific attacks or conditions.
4. These systems also generate an alert based on the presence of the unusual activity.
5. These systems are also capable of detecting "low and slow" attacks.
6. These systems look for individual elements which may be part of an intrusion without waiting for the
7. Completion of an entire sequence of a particular activity.

Disadvantages of SABIDS [8] [9] [10]

1. Statistical methods need accurate statistical distributions, but, not all behaviors can be modeled using purely statistical methods.
2. A majority of the statistical anomaly detection techniques require the assumption of a quasistationary process, which cannot be assumed for most data processed by anomaly detection systems.
3. Its learning process takes days or weeks to become accurate and effective.
4. Setting the threshold is another major problem with SABIDS. Setting the threshold too high will not alert on the necessary traffic, while setting it too low will produce an overabundance of false positives.
5. It generates an unacceptable number of false alarms as it is not able to adapt to legitimate changes in a user's behavior.

a. Operational Model or Threshold Metric

This model is based on the operational assumption that an anomaly can be identified through a comparison of an observation with a predefined limit. Based on the cardinality of observation that happens over a period of time an alarm is raised. The operational model is most applicable to metrics where experience has shown that certain values are frequently linked with intrusions. For an example an event counter for the number of password failures during a brief period, where more than say 10, suggest an failed log-in [3] [7].

In [11], an Adaptive Threshold Algorithm which has been used is based on this sub model. This is a simple and straightforward algorithm that tests whether the traffic measurement, number of SYN packets exceeds over a particular threshold over a given interval. The value of this threshold is set adaptively based on an estimate of the mean number of SYN packets which is computed from recent traffic measurements.

Advantages of Operational model or Threshold Metric: This sub model is a good choice in two cases: [7]

1. When there are not any significant changes in normal behavior.
2. And when the tolerant level of a particular event is known in advance.

The above two advantages help in deciding the boundary values as the data changes its behavior.

Disadvantages of Operational model or Threshold Metric: [7]

1. This approach is not suitable for the malicious activities that may not rely only on one event but more than one event.
2. If the upper and lower boundaries are not significant this sub model will not be able to detect anomalies efficiently.

This sub model cannot be applied to cases where there can be periodic changes in the normal behavior of the data.

b. Markov Process Model or Marker Model

The Markovian Model is used with the event counter metric to determine the normalcy of a particular event, based on the events which preceded it. The model characterizes each observation as a specific state and utilizes a state transition matrix to determine if the probability of the event is high (normal) based on the preceding events. This model is particularly useful when the sequence of activities is particularly important [12]. This model is basically used in two main approaches: Markov chains and hidden Markov models [13].A Markov chain keeps track of an intrusion by examining the system at fixed intervals and maintains the record of its state. If the state change takes place it computes the probability for that state at a given time interval. If this probability is low at that time interval then that event is considered as an anomalous [7].

Advantages of Markov Process Model:

1. This model might be useful for looking at transitions between certain commands where command sequences were important.

2. It can easily accommodate add-on virtual "users" to the community.

Disadvantages of Markov Process Model:

1. A relatively large computing requirement is needed to build the user profile.
2. It cannot handle any drastic change in the normal usage pattern. Hence this model cannot be applied in case of time critical scenarios and where there can be drastic changes in the usage patterns.

c. Statistical Moments or Mean and Standard Deviation Model

The Mean and Standard Deviation Model is based on the traditional statistical determination of the normalcy of an observation based on its position relative to a specified confidence range [12]. In this sub-model, event that falls outside the set interval will be declared as an anomalous [7].

Advantages of Statistical Moments or Mean and Standard Deviation Model: [7][12]

1. This model has the advantage that it "learns" a user's behavior over time instead of requiring prior knowledge of the users activities.
2. It does not require any prior knowledge about normal activity in order to set limits, indeed it learns the normal activities from its observations.
3. It provides more flexibility than threshold model.

Disadvantages of Statistical Moments or Mean and Standard Deviation Model:

It is more complex than other models.

d. Multivariate Model

This model can be applied to intrusion detection for monitoring and detecting anomalies of a process in an information system. This model is similar to the mean and standard deviation model except that it is based on correlations among two or more features. This model would be useful in the situation where two or more features are related [7]. This model permits the identification of potential anomalies where the complexity of the situation requires the comparison of multiple parameters [12]. Many researchers have proposed tests based on multivariate model, for e.g. chi-square (X^2) statistic, Hotel ling's (T^2) test [14] [15]. Multivariate process control techniques like multivariate cumulative sum (MCUSUM) and multivariate exponentially weighted moving average (MEWMA) are typically used to monitor and detecting anomalies of a process in a manufacturing system [16][17].

Advantages of Multivariate Model:

1. This model when applied with chi-square statistic gave better performance in terms of low false alarm rate and a high detection rate.
2. The intrusive events are detected at a very early stage.
3. Misses rates of multivariate techniques are much less than those of univariate.

Disadvantages of Multivariate Model:

1. The computationally intensive procedure of these multivariate process control techniques cannot meet the demands of ID, as ID must deal with large volumes of high-dimensional process data due to a large number (e.g. hundreds or thousands) of behavior measures and a high frequency of event occurrence.
2. Intrusion detection requires a minimum delay of processing each event in an information system to ensure an early indication and warning of intrusions.

e. Time Series Model

The Time Series Model, attempts to identify anomalies by reviewing the order and time interval of activities on the network. If the probability of the occurrence of an observation is low, then the event is labeled as abnormal. This model provides the ability to evolve over time based on the activities of the users [12]. Anomalies in time series data are data points that significantly deviate from the normal pattern of the data sequence.

Advantages of Time series model:

1. This model can be a good approach in the scenarios where future forecast is required based on the present findings.
2. These models can measure trends of behavior over time and detecting gradual but significant shifts in behavior. Hence it can be a good choice where attacks are launched in the form of series.

Disadvantages of Time series model:

1. This model is costly than mean and standard deviation model.
1. This model cannot work efficiently sudden changes in the normal behavior occurs due to anomaly. Hence this model will not be able to detect, efficiently.

Guidelines to select the SABIDS model [7]

1. When dealing with huge amount of network data that might change its behavior over time or the incoming data might keep changing its behavior use of "time series model" can produce better results.
2. In case of distributed attacks "time series model" can prove to be a very promising technique.
3. "Time series model" can be a good choice where attacks are launched in the form of series.
4. While dealing with huge amount of network data but with less security issue and less computational resources available, "mean and standard deviation" is a good choice.
5. When there is requirement of less security, "operational model" is a good approach.
6. When sequences of events are of importance requiring high computational resources, with comparatively high security level, "Markov model" is appropriate.
7. When there are enough resources for computations and the security level is also high then "Multivariate models" are a good choice since they produce better results with less false alarm rate as compared to mean and standard deviation model.

IV. DATA MINING BASED APPROACH

As IDS can only detect known attacks, but it cannot detect insider attacks, the better solution for an IDS can be Data Mining at its core is "pattern finding" and is defined as "the process of extracting useful and previously unnoticed models or patterns from large data stores". The data-mining process tends to reduce the amount of data that must be retained for historical comparisons of network activity, creating data that is more meaningful to anomaly detection [18] [19].
Advantages of data mining approach to IDS:

1. Remove normal activity from alarm data to allow analysts to focus on real attacks.
2. Identify false alarm generators and "bad" sensor signatures.
3. Find anomalous activity that uncovers a real attack.
4. Identify long, ongoing patterns (different IP address, same activity).

Data Mining based approach can further be classified into following techniques:

a. Clustering (unsupervised) of the data into various categories.
b. Association rule discovery: finding normal activity and enabling the discovery of anomalies e.g. ADAM.
c. Classification :(supervised) predicting the category to which a particular instance belongs.

a. Clustering

Clustering is an unsupervised technique for finding patterns in unlabeled data with many dimensions (number of attributes). Mostly k-means clustering is used to find natural groupings of similar instances. Records that are far from any of these clusters indicate unusual activity that may be part of a new attack. The various clustering approaches are Density-based methods, Grid-based methods, Model-based methods, Partitioning methods and Hierarchy methods [20].

Advantages of Clustering-Based Techniques: [21]

1. Clustering-based techniques can operate in an unsupervised mode.
2. Such techniques can often be adapted to other complex data types by simply plugging in a clustering algorithm that can handle the particular data type.
3. The testing phase for clustering-based techniques is fast since the number of clusters against which every test instance needs to be compared is a small constant.

Disadvantages of Clustering-Based Techniques: [21]

1. Performance of clustering-based techniques is highly dependent on the effectiveness of clustering algorithms in capturing the cluster structure of normal instances.
2. Many techniques detect anomalies as a byproduct of clustering, and hence are not optimized for anomaly detection.
3. Several clustering algorithms force every instance to be assigned to some cluster.
4. Several clustering-based techniques are effective only when the anomalies do not form significant clusters among themselves.

b. Association rule discovery

Association Rule mining though a very popular technique usually is very slow, its being replaced by other powerful techniques like clustering and classification. Association rules mining finds correlation between the attributes. The concept of association rule mining for intrusion detection was introduced by Lee, et al. in [22], and is extended in [23, 24, and 25]. This technique was initially applied to the so-called market basket analysis, which aims at finding regularities in shopping behavior of customers of supermarkets. In particular, these kinds of rules, which are called Boolean association rules, try to identify sets of products (items) that are frequently bought together in a transaction. A discovered rule can tell, for example, people who buy butter and milk will also buy bread. Given a set of items $I = \{I_1, I_2, \ldots, I_m\}$ and a database of transactions $D = \{t_1, t_2, \ldots, t_n\}$ where $t_i = \{I_{i1}, I_{i2}, \ldots, I_{ik}\}$ and $I_{ij} \in I$, an association rule is an implication of the form $X \rightarrow Y$ where $X, Y \subset I$ are sets of items called item sets and $X \cap Y = \emptyset$. The support for an association rule $X \rightarrow Y$ is the percentage of transactions in the database that contain $X \cup Y$. The confidence or strength for an association rule $X \rightarrow Y$ is the ratio of the number of transactions that contain $X \cup Y$ to the number of transactions that contain X [26].

Disadvantages of Association rule discovery technique:

1. The execution time or association rule approach increases exponentially with respect to time as the number of attributes increases [27].
2. There is vast number of rules, it is not possible to process all rules in turn.

c. Classification

Intrusion detection can be thought of as a classification problem: we wish to classify each instance as normal or a particular kind of intrusion (attack). Classification is one of the main techniques used in data mining. Its main goal is to learn from class-labeled training instances for predicting classes of new or previously unseen data. : Instances in the training set have labels. New data is classified based on the training set. Basically a classification tree (also called as decision tree) is built to predict the category to which a particular instance belongs [28]. Two methods for building tree are top-down tree and bottom up approach. ID3 and C4.5, two commonly algorithms of decision tree, here the tree is constructed in top-down manner.

Advantages of Classification-Based Techniques: [21]

1. Classification-based techniques, especially the multi-class techniques, can make use of powerful algorithms that can distinguish between instances belonging to different classes.
2. The testing phase of classification-based techniques is fast, since each test instance needs to be compared against the pre computed model.

Disadvantages of Classification-Based techniques:

1. Classification-based techniques rely on the availability of accurate labels for various normal classes, which is often not possible.
2. These techniques assign a label to each test instance, which can also become a disadvantage when a meaningful anomaly score is desired for the test instances.

V. KNOWLEDGE BASED DETECTION TECHNIQUE

Knowledge based detection Technique can be used for both signature based IDS as well as anomaly based IDS. It accumulates the knowledge about specific attacks and system vulnerabilities. It uses this knowledge to exploit the attacks and vulnerabilities to generate the alarm. Any other event that is not recognized as an attack is accepted. Therefore the accuracy of knowledge based intrusion detection systems is considered good. However their completeness requires that their knowledge of attacks be updated regularly [29].

Advantages of Knowledge based detection Technique:

1. The accuracy of this technique is good.
2. It has a very low false alarm rates.
3. It is robust, flexible and scalable. [13]
4. The knowledge gathering is done in detail which makes easier for security officer to take preventive or corrective action.

Disadvantages of Knowledge based detection Technique:

1. The completeness of this technique requires that their knowledge of attacks be updated regularly.
2. It is difficult and time consuming task as maintenance of the knowledge base requires careful and detailed analysis of each vulnerability.
3. These approaches have to face the generalization issue.

Knowledge based detection Technique can further be classified as:

a. State Transition Analysis
b. Expert Systems
c. Signature Analysis
d. Petri Nets

a. State Transition Analysis

State transition analysis, a technique proposed by Porras and Kemmerer [30] was implemented first in UNIX [31] and later in other environments. The technique is conceptually identical to model based reasoning, it describes the attacks with a set of goals and transitions, and represents them as state transition diagrams. State transition diagram is a graphical representation of the actions performed by an intruder to archive a system compromise. In state transition analysis, an intrusion is viewed as a sequence of actions performed by an intruder that leads from some initial state on a computer system to a target compromised state. State transition analysis diagrams identify the requirements and the compromise of the penetration. They also list the key actions that have to occur for the successful completion of an intrusion [32].

b. Expert System

Expert Systems are used primarily by knowledge-based intrusion-detection techniques. The expert system contains a set of rules that describe attacks. Audit events are then translated into facts carrying their semantic signification in the expert system, and the inference engine draws conclusions using these rules and facts. This method increases the abstraction level of the audit data by attaching semantic to it. [33] It also encodes knowledge about past intrusions, known system vulnerabilities and the security policy. As information is gathered, the expert system determines whether any rules have been satisfied [34]. Expert systems can be used for both signature based IDS as well as anomaly based IDS. Following are two examples that have used this approach:-

1. Wisdom and Sense [35] is an intrusion detection tool that detects statistical anomalies in the behavior of users. The tool first builds a set of rules that statistically describe the behavior of the users based on recordings of their activities over a given period of time. Current activity is then matched against these rules to detect inconsistent behavior. The rule base is rebuilt regularly to accommodate new usage patterns.
2. AT and T's Computer Watch [36] is a tool delivered with AT and T's UNIX/MLS multilevel security operating system. This tool checks the actions of users according to a set of rules that describe proper usage policy, and flags any action that does not fit the acceptable patterns. This approach is useful for policy based usage profiles, but is less efficient than the statistical approach for processing large amounts of audit information.

c. Signature Analysis

Signature analysis follows exactly the same knowledge-acquisition approach as expert systems, but the knowledge acquired is exploited in a different way. The semantic description of the attacks is transformed into information that can be found in the audit trail in a straightforward way. For example, attack scenarios might be translated into the sequences of audit events they generate, or into patterns of data that can be sought in the audit trail generated by the system. This method decreases the semantic level of the attacks description. This technique allows a very efficient implementation and is therefore applied in various commercial intrusion-detection products e.g. Haystack [37].

The main disadvantage of this technique is that like all knowledge-based approaches there is the need for frequent updates to keep up with the stream of new vulnerabilities discovered.

d. Petri Nets

To represent signatures of intrusions, IDIOT [38], a knowledge-based intrusion-detection system developed at Purdue University, uses Colored Petri Nets (CPN). The advantages of CPNs are their generality, their conceptual simplicity, and their graphical representability. System administrators are assisted in writing their own signatures of attacks and integrating them in IDIOT. Owing to the generality of CPNs, quite complex signatures can be written easily. However, matching a complex signature against the audit trail may become computationally very expensive [37].

VI. MACHINE LEARNING BASED DETECTION TECHNIQUE

Machine learning can be defined as the ability of a program and/or a system to learn and improve their performance on a certain task or group of tasks over time. Machine learning techniques focus on building a system that improves its performance based on previous results i.e. machine learning techniques have the ability to change their execution strategy on the basis of newly acquired

information [5]. This feature could make it desirable to use in all situations, but the major drawback is their resource expensive nature. In many cases, the machine learning technique coincides with that of the statistical techniques and data mining techniques. [13]
This technique can further classified as:

a. Bayesian Approach
b. Neural Networks
c. Fuzzy Logic
d. Genetic Algorithms
e. Support vector machines

a. Bayesian Approach

A Bayesian Approach is a graphical model that encodes probabilistic relationships among variables of interest. It is a popular representation for encoding uncertain expert knowledge in expert systems. More recently, researchers have developed methods for learning Bayesian networks from data. The techniques that have been developed are new and still evolving, but they have been shown to be remarkably effective for some data-analysis problems. Bayesian Approach is generally used for intrusion detection in combination with statistical techniques, yielding several advantages [39], as shown below. However, as pointed out in [40], a serious disadvantage of using Bayesian networks is that their results are similar to those derived from threshold-based systems, while considerably higher computational effort is required [41] proposed a method based on a technique called Pseudo-Bayes estimators to enhance an ABIDS's ability to detect new attacks while reducing the false alarm rate as much as possible.

Advantages of Bayesian Approach: [39]

1. It readily handles situations where some data entries are missing (incomplete data sets).
2. It allows one to learn about causal relationships.
3. As the model has both a causal and probabilistic semantics, it is an ideal representation for combining prior knowledge (which often comes in causal form) and data.
4. It provides an efficient and principled approach for avoiding the over fitting of data.

b. Neural Networks

In the neural network approach the systems learn to predict the next command based on a sequence of previous commands by a specific user. Neural networks provide a solution to the problem of modeling the users' behavior in anomaly detection because they do not require any explicit user model. Neural networks for intrusion detection were first introduced as an alternative to statistical techniques in the IDES intrusion detection expert system [42]. Ghosh et al. [43] found that a "well trained, pure feed forward, back propagation neural network" performed comparably to a basic signature matching system. Various neural networks can be used for anomaly based IDS like Multi layered Perceptrons, Radial Basis Function-Based, Hopfield Networks etc.
For building a neural network an IDS consists of three phases: [44]
1. The collection of training data by obtaining the audit logs for each user for a certain period. A vector is formed for each day and each user, which shows how often the user, executed each command.
2. Train the neural network to identify the user based on the command distribution vectors.
3. The neural network to identify the user based on the command distribution vector. If the network's suggestion is different from the actual user, an anomaly is signaled.

c. Fuzzy Logic Approach

Fuzzy logic techniques have been in use in the area of computer and network security since the late 1990's [45]. Fuzzy logic plays an important role in implementing ABIDS. The fuzzy logic part of the system is mainly responsible for both handling the large number of input parameters and dealing with the inexactness of the input data. When combined with data mining, it reduces the size of the input data sets and selects features that highlight anomalies; fuzzy logic can be an effective means of defining network attacks. Dickerson et al. developed the Fuzzy Intrusion Recognition Engine (FIRE) using fuzzy sets and fuzzy rules [19].

A fuzzy logic technique has been used in correlation with Intrusion Detection because of its following characteristics: [19] [46] [47]

1. Various quantitative parameters used for Intrusion Detection e.g., CPU usage time, activity frequency, connection interval, etc., are fuzzy in nature [49].
2. The concept of security itself is fuzzy as stated by Bridges et al. [49].
3. Fuzzy systems can readily combine inputs from varying sources [50].
4. The degree of alert that can occur with intrusion is often fuzzy [50].
5. Fuzzy rules allow us to easily construct if-then rules that help in describing security attacks [22].

d. Genetic Algorithms

Genetic algorithms were originally introduced in the field of computational biology. It uses the computer to implement the natural selection and evolution. This concept comes from the ‘‘adaptive survival in natural organisms”. The algorithm starts by randomly generating a large population of candidate programs. Some type of fitness measure to evaluate the performance of each individual in a population is used. A large number of iterations is then performed that low performing programs are replaced by genetic recombination of high performing programs. That is, a program with a low fitness measure is deleted and does not survive for the next computer iteration [48]. Researchers have tried to integrate it with IDSs: Crosbie and Spafford [49] have used a genetic algorithm for sparse trees to detect anomalies. They attempted to minimize the occurrence of false positives by utilizing human input in a feedback loop.

e. Support vector machines

Support vector machines (SVM) is proposed by Vapnik in 1998. SVM first maps the input vector into a higher dimensional feature space and then obtain the optimal separating hyper-plane in the higher dimensional feature space. Moreover, a decision boundary, i.e. the separating hyper-plane, is determined by support vectors rather than the whole training samples and thus is extremely robust to outliers. In particular, an SVM classifier is designed for binary classification. That is, to separate a set of training vectors which belong to two different classes? Note that the support vectors are the training samples close to a decision boundary. The SVM also provides a user specified parameter called penalty factor. It allows users to make a tradeoff between the number of misclassified samples and the width of a decision boundary. Eskin et al. [50] and Honig et al. [51] used an SVM in addition to their clustering methods for unsupervised learning. The achieved performance was comparable to or better than both of their clustering methods. Mukkamala, Sung, et al. [52] [53] used a more conventional SVM approach. They used five SVMs, one to identify normal traffic, and one to identify each of the four types of malicious activity in the KDD Cup dataset. Every SVM performed with better than 99% accuracy, even using seven different variations of the feature set. As the best accuracy they could achieve with a neural network (with a much longer training time) was 87.07%, they concluded that SVMs are superior to neural nets in both accuracy and speed.

VII. EXAMPLES OF SOME OF THE PROJECTS OF ABIDS

FIRE (Fuzzy Intrusion Recognition Engine)

It is an ABIDS that uses fuzzy system to access malicious activity against computer networks. FIRE uses simple data mining techniques to process the network input data and generate fuzzy sets for every observed feature. The fuzzy sets are then used to define fuzzy rules to detect individual attacks. FIRE does have any sort of model representing the current state of the system, but it relies on attack specific rules for detection. It creates and applies fuzzy logic rules to the audit data to classify it as normal or anomalous [9]. The important features of FIRE are as follows: [19]

1. It shows that fuzzy systems can be used as an intrusion detection method.
2. It has identified data sources that are the best inputs to the fuzzy intrusion detection system.
3. It has given the best methods for representing network input data.
4. It has shown how the system can be scaled to distributed intrusion detection involving multiple hosts and/or networks.
5. It use readily available software and hardware as much as possible,
6. And it is a tool that can be accessed by system/security administrator.

Advantages of FIRE:

1. It does not rely on the entire model for representing current state of the system but it relies on attack specific rules for detection.
2. It creates and applies the concept of fuzziness to the input data to classify it as normal or anomalous.
3. This approach is particularly effective against port scans and probes attacks.

Disadvantages of FIRE:

1. The primary disadvantage to this approach is the rule generation process is very laborious.

ADAM (Audit Data Analysis and Mining)

ABIDS has the ability to detect new attacks, but in practice this is far from easy. Anomaly detection has the potential to generate too many false alarms, and it is very time consuming and labor expensive to sift true intrusions from the false alarms [41]. In order to improve the detection and false alarm rates ADAM was designed [54]. It was developed at the Center for Secure Information Systems of George Mason University. It is an real-time anomaly detection system, which uses a module to classify the suspicious events into false alarms or real attacks. It works in two phases: training phase and detection phase [7]. It builds a profile of normal activities over attack-free training data, and then detects attacks with the previously built profile. It is essentially a testbed for using data mining

techniques to detect intrusions. ADAM uses a combination of association rules mining and classification to discover attacks in a TCPdump audit trail [55]. ADAM is unique in two ways. First, ADAM uses data mining to build a customizable profile of rules of normal behavior, and a classifier that sifts the suspicious activities, classifying them into known attack, unknown attacks or false alarms. Secondly, ADAM is designed to be used on-line (in real time), a characteristic achieved by using incremental mining algorithms that use a sliding window of time to find suspicious events [56].

Advantages of ADAM:

1. It has the ability to work in real time.
2. It is able to detect novel attacks (unknown attacks).
3. When used with classifiers the number of false alarms is greatly reduced as the abnormal associations that belong to normal instances are filtered out. [56]

Disadvantages of ADAM

1. It has high dependency on training data for normal activities.

We are very confident that ADAM will become, with time, a very strong tool to help security officers in their daily work against intruders. [56]

Haystack

Haystack is a prototype system for the detection of intrusions in multi-user Air Force computer systems, which mainly runs on a Unisys (Sperry) 1100/60 mainframe and the OS/1100 operating system (the standard Air Force computing platform at that time) [57]. Haystack reduces voluminous system audit trails to short summaries to help the System Security Officer (SSO) detect and investigate intrusions, particularly by insiders (authorized users) [58]. It is one of the earliest examples of a statistical anomaly-based intrusion detection system. It defines a range of values that were considered normal for each feature. If during a session, a feature fell outside the normal range, it keeps a score. It raises an alarm if this score becomes too large. It also maintained a database of user groups and individual profiles. If a user had not previously been detected, a new user profile with minimal capabilities was created using restrictions based on the user's group membership [2].

Advantages of Haystack

1. It provides the operational utility for SSO to reduce voluminous system audit trails to short summaries for the further investigation of potential computer intrusions.
2. It allows the SSO to new reports by making the "Ad-hoc" queries against the database.

Disadvantages of Haystack

1. It was designed to be used offline, so it failed to be used for real-time intrusion detection systems since doing so required high-performance systems.
2. Because of its dependence on maintaining profiles, a common problem for system administrators was the determination of what attributes were good indicators of intrusive activity.

VIII. CONCLUSION

Intrusion detection is currently attracting interest from both the research community and commercial companies. In this paper, we have given an overview of the current state-of-the-art of ABIDS, based on a proposed taxonomy illustrated with examples of past and current projects. This taxonomy also highlights the properties of ABIDS and covers the past and current developments adequately. Each of its technique has its own advantages and disadvantages. We believe that no single criterion can be used to completely defend against computer network intrusion. There is no single version of it that can be used as a standard solution against all possible attacks. It is both technically difficult and economically costly to build and maintain computer systems and networks that are not susceptible to attacks. The technique to be selected depends on the specifications of the type of anomalies that the system is supposed to face, the type and behavior of the data, the environment in which the system is working, the cost and computation limitations and the security level required.

REFERENCES

[1] Anderson, J.P ,"Computer Security Threat Monitoring and Surveillance", Technical Report, J.P. Anderson Company, Fort Washington, Pennsylvania, April 1980.

[2] D. E. Denning, P. G. Neumann, "Requirements and model for IDES-A real-time intrusion detection system", Computer Sci. Lab, SRI International, Menlo Park, CA, Tech. Rep., 1985.

[3] Dorothy E. Denning, "An Intrusion-Detection Model", IEEE Transactions on Software Engineering, Vol. SE-13, No. 2, pp. 222-232, Feb 1987.

[4] Parker, D.B., "Demonstrating the elements of information security with treats", In Proceeding of the 17th National Computer Security Conference, pp. 421-430, 1994.

[5] Animesh Patcha, Jung-Min Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends", Elsevier, Science Direct, Computer Networks, 51, pp. 3448-3470, 2007.

[6] Markos Markou, Sameer Singh, "Novelty Detection; a review-part 2: Neural Network based approaches", Signal Processing, Vol. 83, pp. 2499-2521, 2003.

[7] A. Qayyum, M. H. Islam, and M. Jamil, "Taxonomy of Statistical Based Anomaly Detection Techniques for Intrusion Detection", IEEE, International Conference on Emerging Technologies, Islamabad, pp. 270-276, September 17-18, 2005.

[8] statistical-based-intrusion-detection www.symantec.com/connect/.../statistical-based-intrusion-detection

[9] K.Parvathi Devi and Y.A Siva Prasad, "Study of Anomaly Identification Techniques in Large Scale Systems", International Journal of Computer Trends and Technology, Vol.3, Issue1, 2012.

[10]James Cannady and Jay Harrel, "A Comparative Analysis of Current Intrusion Detection Technologies", Proceedings of Technology in Information Security Conference (TISC), pp. 212-218, 1996.

[11] Vasilios A. Siris, and Fotini Papagalou, "Application of anomaly detection algorithms for detecting SYN flooding attacks", Elsevier, Computer Communications, Vol. 29, pp. 1433–1442, 2006.

[12] James Cannady Jay Harrell, "A Comparative Analysis of Current Intrusion Detection Technologies", Proceeding of 4th Technology for Information Security Conference, TISC'96, Houston, TX, May 1996.

[13] P. Garcı´a-Teodoroa, J. Dı´az-Verdejo, G. Macia´-Ferna´ndez and E. Va´zquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges" , Elsevier, Computers and Security, Vol. 28 ,pp. 18-28, 2009.

[14] Nong Ye, Syed Masum Emran, Qiang Chen, and Sean Vilbert, "Multivariate Statistical Analysis of Audit Trails for Host-Based Intrusion Detection" , IEEE Transaction on Computers, Vol. 51, No. 7, July 2002.

[15] Nong Ye and Qiang Chen, "An Anomaly Detection Technique Based on a Chi-Square Statistic for Detecting Intrusions into Information Systems", Quality and reliability Engineering International, Vol.17, No.2, pp. 105-112, 2001.

[16]Woodall WH,and Ncube MM, "Multivariate CUSUM quality control procedures", Techno metrics, Vol. 27, pp. 185–192, 1985.

[17] Lowry CA, Woodall WH, Champ CW, and Rigdon SE, "Multivariate exponentially weighted moving average control chart" , Techno metrics ,Vol. 34, pp. 46–53, 1992.

[18] M. Sathya Narayana, B. V. V. S. Prasad, A. Srividhya and K. Pandu Ranga Reddy, "Data Mining Machine Learning Techniques – A Study on Abnormal Anomaly Detection System", International Journal of Computer Science and Telecommunications ,Vol. 2, Issue 6, September 2011.

[19] J. E. Dickerson and J. A. Dickerson, "Fuzzy network profiling for intrusion detection" , In 19th International Conference of the North American Fuzzy Information Processing Society (NAFIPS), Atlanta, GA, pp. 301 – 306 , 2000 .

[20] Jian Pei, Shambhu, J. Upadhyaya, Faisal Farooq and Venugopal Govindaraju , "Data Mining for Intrusion Detection – Techniques, Applications and Systems" , Data Mining Techniques for Intrusion Detection and Computer Security, University at Buffalo, New York.

[21] Varun Chandola, Arindam Banerjee, and Vipin Kumar, "Anomaly Detection: A Survey", ACM Computing Surveys, Vol. 41, No. 3, Article 15, July 2009.

[22] Lee, W., Stolfo, S., "Data Mining Approaches for Intrusion Detection", Proceedings of the 7th USENIX Security Symposium, pp. 79-94, 1998.

[23] Barbara D, Couto J, Jajodia S, and Wu N, "ADAM: A Testbed for Exploring the Use of Data Mining in Intrusion Detection", SIGMOD Record, Vol. 30, No. 4, pp. 15-24, 2001.

[24] Lee, W. Stolfo and S. Kui, M., "A Data Mining Framework for Building Intrusion Detection Models", IEEE Symposium on Security and Privacy, pp.120-132, 1999.

[25] Manganaris S, Christensen M, Zerkle D, and Hermiz K, "A Data Mining Analysis of RTID Alarms", Proceedings of Recent Advances in Intrusion Detection, Second International Workshop, 1999.

[26] Margaret H. Dunham, "Data Mining Introductory and Advanced Topics", Prentice Hall, 2003, ISBN 0-13-088892-3

[27] Arman Tajbakhsh, Mohammad Rahmati, and Abdolreza Mirzaei, "Intrusion detection using fuzzy association rules", Applied Soft Computing, Vol. 9, pp. 462–469, 2009.

[28] Mohammadreza Ektefa, Sara Memar, Fatimah Sidi , and Lilly Suriani Affendey, "Intrusion Detection Using Data Mining Techniques Information Retrieval & Knowledge Management", CAMP, 2010.

[29] Herve Debar, Marc Dacier and Andreas Wespi, "Towards a Taxonomy of Intrusion Detection Systems", Elsevier, Computer Networks, Vol. 31, pp. 805–822, 1999.

[30] Phillip A, Porras and Alfonso Valdes "Live traffic analysis of tcp/ip gateways", Proceeding ISOC Symposium on Network and Distributed System Security, San Diego, CA, March1998.

[31] Koral Ilgu, "Ustat: A real- time intrusion detection system for Unix", Proceeding IEEE Symposium on Research in Security and Privacy, Oakland, CA, pp. 16-28, May 1993.

[32] E. Biermann, E.Cloete, and L.M. Venter, "A comparison of Intrusion Detection systems", Elsevier, Computers & Security, Vol. 20, pp. 676-683, 2001.

[33] T. Lunt, and R. Jagannathan, "A prototype real-time intrusion detection expert system", Proceeding of Symposium on Security and Privacy, Oakland, CA, pp. 59-66, April 1988.

[34] Frank, J., "Artificial intelligence and intrusion detection: current and future directions", In Proceedings of the 17th National Computer Security Conference, October 1994.

[35] H. S. Vaccaro and G. E. Liepins, "Detection of anomalous computer session activity", Proceeding IEEE Symposium on Research in Security and Privacy, pp. 280-289, 1989.

[36] Cheri Dowell and Paul Ramstedt, "The Computer Watch data reduction tool" ,Proceeding 13th National Computer Security Conference, Washington, DC, pp. 99-108, October 1990.

[37] Herve Debar, Marc Dacier, Andreas Wespi, "Towards a taxonomy of intrusion-detection systems", Elsevier, Computer Networks, Vol. 31, pp. 805–822, 1999.

[38] S. Kumar, and E. Spafford, "A pattern matching model for misuse intrusion detection", Proceeding 17th National Computer Security Conference, pp. 11–21, October 1994.

[39] David Heckerman, "A Tutorial on Learning with Bayesian Networks", Microsoft Research, Technical Report MSRTR 95, 2006.

[40] Kruegel C., Mutz D., Robertson W., Valeur F, "Bayesian event classification for intrusion detection", In: Proceedings of the 19th Annual Computer Security Applications Conference, 2003.

[41] D. Barbara, N. Wu, and S. Jajodia, "Detecting novel network intrusions using bayes estimators", In Proceedings of the First SIAM International Conference on Data Mining , Chicago, USA, Apr. 2001.

[42] Debar H, Becker M, and Siboni D, "A Neural Network Component for an Intrusion Detection System", IEEE Computer Society Symposium on Research in Security and Privacy, Los Alamitos Oakland, CA, pp. 240–250, May 1992.

[43] Ghosh A, K. A Schwartzbard, and M Schatz, "Learning program behavior profiles for intrusion detection", In Proceeding of 1st USENIX, 9-12 April, 1999.

[44] E. Biermann, E.Cloete, and L.M. Venter, "A comparison of Intrusion Detection systems", Elsevier, Computers & Security, Vol. 20, pp. 676-683, 2001.

[45] H. H. Hosmer, "Security is fuzzy!: Applying the Fuzzy Logic Paradigm to the Multipolicy Paradigm", In 1992-1993workshop on New security paradigms, Little Compton, Rhode Island, United States 1993.

[46] S. M. Bridges and R. B. Vaughn, "Fuzzy Data Mining and Genetic Algorithms Applied to Intrusion Detection", National Information Systems Security Conference, Baltimore, MD, 2000.

[47] J. E. Dickerson, J. Juslin , O Koukousoula, and J. A. Dickerson, " Fuzzy Intrusion Detection", IFSA World Congress and 20th North American Fuzzy information Processing Society, International Conference , Vancouver, British Columbia, Vol. 3, pp: 1506-1510, 2001.

[48] Chih-Fong Tsai , Yu-Feng Hsu , Chia-Ying Lin and Wei-Yang Lin , "Intrusion detection by machine learning: A review ", Elsevier, Expert Systems with Applications, Vol. 36, pp. 11994–12000, 2009.

[49] Crosbie, M. and E. H. Spafford, "Active defense of a computer system using autonomous agents", Technical Report CSD-TR-95-008, Purdue University, West Lafayette, 15 February 1995.

[50] Eskin E, A Arnold, M Preraua, L Portnoy, and S. J. Stolfo, "A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data", In D. Barbar and S. Jajodia (Eds.), Data Mining for Security Applications, Boston: Kluwer Academic Publishers, May 2002.

[51] Honig, A., A. Howard, E. Eskin, and S. J. Stolfo, "Adaptive model generation: An architecture for the deployment of data mining based intrusion detection systems", In D. Barbar and S. Jajodia (Eds.), Data Mining for Security Applications. Boston, Kluwer Academic Publishers, May 2002.

[52] Mukkamala S, and A. H. Sung, "Identifying key variables for intrusion detection using neural networks", Proceedings of 15th International Conference on Computer Communications, pp. 1132-1138, 2002.

[53] Mukkamala, S. and A. H. Sung, "Identifying significant features for network forensic analysis using artificial intelligent techniques", International Journal of Digital Evidence, Vol. 1, No. 4, pp. 1-17, 2003.

[54] Peyman Kabiri and Ali A. Ghorbani, "Research on Intrusion Detection and Response: A Survey", International Journal of Network Security, Vol.1, No.2, pp.84–102, Sep. 2005.

[55] Daniel Barbara , Sushil Jajodia, Ningning Wu and B. Speegle, "Mining Unexpected Rules in Network Audit Trails", Technical Report, George Mason University, ISE Dept. September 1999.

[56] Daniel Barbara ,Julia Couto, Sushil Jajodia, Leonard Popyack, Ningning Wu and George Mason, "ADAM: Detecting Intrusions by Data Mining" , Proceedings of the 2001 IEEE Workshop on Information Assurance and Security United States Military Academy, West Point, NY, 5-6 June 2001.

[57] Stefan Axelsson, "Intrusion Detection Systems: A Survey and Taxonomy", Technical Report No 99-15, Dept. of Computer Engineering, Chalmers University of Technology, Sweden, March 2000.

[58] Stephen E. Smaha, "Haystack: An Intrusion Detection System", IEEE, Fourth Aerospace Computer Security Applications Conference, Orlando, FL, pp.37 – 44, 1988.

AUTHORS

1. **Manasi Gyanchandani:** PhD pursuing, M.E & B.E. in Computer Science & Engineering.
E-mail address: manasi_gyanchandani@yahoo.co.in
2. **Dr. J. L.Rana**: PhD from IIT Mumbai MS from USA (Hawaii), Guided 10 PhDs.
E-mail address: jl_rana@yahoo.com
3. **Dr. R.N.Yadav** : PhD from IIT Kanpur, B.E. and Mtech from Motilal Nehru Regional Engineering College Allahabad, and Maulana Azad National Institute of Technology, Bhopal
E-mail address: myadav@gmail.com