Intrusion Detection Using Soft Computing Techniques

Koushal Kumar
Assistant Professor
Department of Computer applications
Sikh National College Qadian (GSP)
(kaushal_kumar302@yahoo.com)

Simranjit Singh
Assistant Professor
Department of Computer Science
Guru Nanak Dev University
(singhs_simran@yahoo.co.in)

Abstract

As the popularity and usage of Internet increases all over the world, security issues have also become very essential that have to be considered. In literature, different techniques from various disciplines have been utilized to develop efficient security methods but those techniques always have some shortcomings. In this chapter the author discuss soft computing techniques that have been applied in Intrusion Detection System (IDS). The scope of this chapter will encompass core techniques of soft computing including artificial neural networks, Fuzzy Logic Genetic algorithms and hybridization of these approaches. In this chapter the research contribution of each of above mentioned techniques will be systematically summarized and compared that will allows us to clearly define existing research challenges, and to highlight promising new research directions. The findings of this chapter should provide useful insights into the current IDS literature and be a good source for anyone who is interested in the application of soft computing approaches to IDSs.

Keywords: Artificial Neural Network (ANN), Genetic Algorithm (GA), Fuzzy Logic (FL), Network Security, Back Propagation Network (BPN)

Introduction:

The Internet continuously growing exponentially with its usage in personal government and business applications which ultimately benefits to end users. Securing the digital assets (information) is a major concern in the present digital information era. Security of our computer systems (Personnel computer) and the networks (wired and wireless) to which our computer is connected becoming a significant topic of concern. Internet was originally designed by keeping functionality but not security in mind [1]. The Internet connects hundreds of millions of computers across the world and each computer running on heterogeneous hardware and software platforms for the purpose of resource sharing. However this interconnectivity among computers also enables malicious users to misuse resources and mount Internet attacks [2]. Security vulnerabilities if left unnoticed can destroy your whole system or network. Vulnerabilities are specific flaws in the hardware or software implementation of your networks or your application that may allow attackers to expose, alter, and change the sensitive information. Vulnerability could be weakness in your technology, configuration of network components or in the security policies you have applied. Without adequate and proper network security mechanisms many individuals, governments and businesses cannot survive for a long time period on the internet. Vulnerabilities can be exploited by hackers and results in devastating consequences and harmful attack [3]. Attack is any attempt to destroy expose, alter, disable, steal, or gain unauthorized access to or make unauthorized use of a resource. The dependency of more and more services on computer technology has resulted in the increase of computer related threats. As the number of users who are using internet is growing with time our networks is facing more and more electronic attacks. The increase in the number and severity of threats to our computer networks
has given birth to a new field of study. Network Security & Cryptography is the field of study dealing with protecting our networks and data transmission over wired/wireless network. Data Security is the main aspect of secure data transmission over unreliable network. Data Security is a challenging issue of data communications today that touches many areas including secure communication channel, strong data encryption technique and trusted third party to maintain the database [4]. Although computer scientists have developed several hi-tech innovations for computer security, but still it is a fact that it is nearly impossible to have a completely secured system which can protect our computer system from every kind and every time from attackers.

**Attack types and their Classification**

Broadly we can classify security related attacks in two categories.

1) Active attack
2) Passive attack

![Figure 1: Types of attacks](image1.png)

The major difference between both kinds of attack is the way how intruder makes use of information. The term passive indicates that the attacker does not attempt to perform any modification to the data. In passive attack mode intruder only monitor the transmission of information that is send by source to destination but does not affect the system resources. Passive attackers mainly perform the following activities
- Network traffic analysis
- Decrypting weakly encrypted message
- Collecting authentication information (passwords)

**Passive attacks** are very hard to detect so general approach to deal with passive attacks is to apply prevention actions. On the other hands **Active attacks** are based on modification of the original secret message or creating a false message so that receiver could not understand the original message. These kinds of attacks can be detected with some extra effort. In active attacks intruder try to access system resources and harm the system and its operations. Active attacks can be of following categories:
- Masquerade
- Replay
- Alteration of message
- Denial of service attack (DOS)

![Figure 2: Categories of active attacks](image2.png)

**Masquerade** is a type of attack where the attackers act as an authorized user in order to gain access of the resources. **Alteration** is a kind of attack which involves alteration of information in an unauthorized manner that is intended to appear genuine to the user. **Denial of Services (DOS)** is a type of attack that prevent legitimate users from accessing some services or resource, which they are eligible for. **Spoofing** this is an attempt by someone or something to masquerade as someone else. This type of attack is usually considered as an access attack. Example IP spoofing.

**Replay** attacks are used for access or modification attacks. This occurs when information is captured over a network [5].

**What is Intrusion Detection System (IDS) and traditional detection approaches:**

**Intrusion** is defined as an act of violating the confidentiality, integrity, or availability of a computer or a
computer network system. **Intrusion detection system** (IDS) is one of most important system being used to detect the Internet attacks which can be either host based or network based. IDS plays a major role in maintaining and keeping information secure in any networking environment. An intrusion detection system records and monitors all inbound and outbound network traffic and identify suspicious activity which tries to break network security. Intrusion Detection System (IDS) can be software or a combination of software and hardware that automatically detects computer intrusions and reacts properly in order to protect computers and networks from activity occurring in your network [6]. Traditionally Intrusion detection approaches are classified into following categories

- Misuse detection
- Anomaly detection
- Specification-based detection

**Misuse detection**: misuse based IDS uses information of already occurred attack patterns to identify attacks. Intrusions are detected by matching actual behavior recorded in audit trails with known suspicious patterns. So from its working it is clear that misuse detection is fully effective in detecting known attacks but it is useless when encountered with unknown or novel forms of attacks for which the signatures are not yet available. Figure 3 shows the typical misuse detection model. Model consists of four major components: namely, Data collection, system profile, misuse detection and response. Data are collected from one or many data sources including audit trails, network traffic, system call trace, etc. Collected data are transferred into a format that is understandable by the other components of the system. The system profile is used to characterize normal and abnormal behaviors. The profiles are matched with actual system activities and reported as intrusions in case of deviations. Four classes of techniques are commonly used to implement misuse detection, namely pattern matching, rule-based techniques, state-based techniques, and data mining.

Modify/Add existing rules

![Figure 3: A typical misuse detection model](image)

**Table I: Summary of various IDS**

<table>
<thead>
<tr>
<th>Name of system</th>
<th>Processing Criteria</th>
<th>Source of audit data</th>
<th>Type of Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSM (1990)</td>
<td>Hybrid</td>
<td>N/w</td>
<td>Passive</td>
</tr>
<tr>
<td>Bro (1998)</td>
<td>Signature</td>
<td>N/w</td>
<td>Passive</td>
</tr>
<tr>
<td>MIDAS (1988)</td>
<td>Hybrid</td>
<td>Host</td>
<td>Passive</td>
</tr>
<tr>
<td>Haystack (1988)</td>
<td>Hybrid</td>
<td>Host</td>
<td>Passive</td>
</tr>
<tr>
<td>IDES (1992)</td>
<td>Anomaly</td>
<td>Host</td>
<td>Passive</td>
</tr>
<tr>
<td>Comp Watch (1990)</td>
<td>Anomaly</td>
<td>Host</td>
<td>Passive</td>
</tr>
<tr>
<td>ASAX (1992)</td>
<td>Signature</td>
<td>Host</td>
<td>Passive</td>
</tr>
</tbody>
</table>
Anomaly Based: Anomaly based IDS uses normal instances as the base data to operate on. Any instance or behavior deviating from this normal behavior is termed anomalous and is categorized as an attack. Anomaly detection studies start by forming an opinion on what the normal attributes for the observed objects are, and then decide what kinds of activities should be flagged as intrusions and how to make such particular decisions. A typical anomaly based model is shown in Figure 4. It consists of four components, namely data collection, normal system profile, anomaly detection and response. Normal user activities or traffic data are obtained and saved by the data collection component. The anomaly detection component decides how far the current activities deviate from the normal system profiles and what percentage of these activities should be flagged as abnormal.

Generate new profile significant deviations?

Specification-based detection: Specification based approach is different from misuse and anomaly based techniques. Instead of learning system behaviors in specification-based systems the experts’ knowledge determines the operating limits of a system. Once the correct (or allowed) system behavior is specified. The events deviating from the specification would generate an alert [7].

What are the shortcomings of conventional intrusion detection schemes?

In the last few years, different techniques were proposed and successfully deployed to secure the computer systems from unauthorized use. Misuse detection and Anomaly detection both have their own advantages and disadvantages. Although a misuse detection approach is effective and efficient in detecting known attacks, it rarely detects new attacks. On the other hand, an anomaly detection approach is very good in detecting unknown attacks; however, it may generate a high number of false alarms because it can report unknown normal behavior as anomalous [8]. In terms of hardware and software these techniques needs extra software hardware or combination of both like firewalls secure network protocols, password protection user authentication data encryption the avoidance of programming errors and many others.
security tools. If a password is weak and is compromised, user authentication cannot prevent unauthorized use. Firewalls are vulnerable to attack and prone to errors in case of wrong configuration or ambiguous security policies. Similarly programming errors cannot be avoided as the complexity of the system and applications software is changing rapidly [9]. Some of major performance issues in traditional IDS were following

- Speed of data processing
- More memory requirements
- Less accuracy in detection

**Soft Computing Techniques In Intrusion detection system:**

In order to tackle the difficulties of conventional intrusion detection approaches, different soft computing techniques have been applied to the intrusion detection problem. **Soft Computing** is an innovative and emerging approach to deliver computationally intelligent systems. In opposite with conventional artificial intelligence techniques which only deal with precision, certainty, and work in time constraint scenario on the other hand the soft computing deals with imprecision, uncertainty, partial truth, and approximation to achieve practicability robustness and low solution cost [10]. The principal constituents of Soft Computing are Fuzzy Logic, Neural Computing, and Evolutionary Computation. Empirical results from literature clearly show that soft computing approach could play a major role for Intrusion detection. In the next section we will discuss major constituents of soft computing one by one with their effectiveness in the field of Intrusion detection.

**Artificial neural network:** The major motivation to adopt the artificial neural networks in many applications filed is due to its inception that human brain computes an entirely different way from the conventional digital computer (Haykin, 1999, P.1). The main driving force which provides neural networks abundant computational power is its massive interconnection structure among its neurons. Neural networks have already been used to solve many problems related to pattern recognition, data mining, data compression and research is still underway with regards to intrusion detection systems. Artificial Neural Networks have been applied to solve the Intrusion detection problem since the early 90s era. Artificial Neural networks roughly based upon the organizational structure of the human brain and have ability to acquire knowledge through learning and store it in inter-neuron connections known as synaptic weights. The performance of neural networks depends on the architecture, algorithms and learning model chosen to collect and process data [12].

**Architecture:** ANN architecture can be a Single Layer feed forward, multiple layer feed-forward, Recurrent, Back propagation neural network etc. Single layer networks have only one layer of neurons connected individually to input points while multiple layers usually have several layers of neurons to process the data. In a single feed forward network the information move forward from input layer to output layer without backward feedback. Multiple layer models use algorithms such as back propagation to learn; output values are compared with the result values in order to correct errors. The acquired information is then forwarded back to the network for self correction. Recurrent networks use multiple layers and back propagation for learning.

**Learning algorithms:** There are a variety of algorithms used for learning including: error correction learning, Hebbian learning, competitive learning, self organizing maps, back propagation ((Hagan at al. 1996), snap-drift algorithm neo cognition, feature map, competitive learning, adaptive resonance theory, principal component, perceptron, decision-based, multilayer perceptron, temporal dynamic model, hidden Markov model, Hamming net, Hopfield net, combinatorial optimization etc.

**Learning model:** Model can be Supervised or Unsupervised. Supervised models have been the mainstream of neural development for some time. The training data consist of many pairs of input/output training patterns and the learning process relies on assistance (Kung,1993).While in the learning phase the neural network learn the desired output for a given input. Multiple layer perceptron (MLP) algorithm is used often with supervised models. In the case of unsupervised models, the network gain knowledge without specifying the required output during the learning phase. The self-organizing map (SOM) algorithm is associated frequently with unsupervised models [13].
Artificial neural network in intrusion detection:
A review of available literature shows that intrusion detection approaches that had been used in the last few years has some major issues
Like high false alarm rate, failure to detect new intrusions, and lack of supervised learning. Neural networks are used in intrusion detection as classifiers (misuse detection) as well as for clustering (anomaly detection), giving good results as compare to traditional approaches.

In the last years, several Artificial Neural Network (ANN) techniques such as multilayer perceptron network (MLPN)/Back Propagation network (BPN), Radial basis function (RBF), Recurrent neural networks(RNN), have been applied in many IDS models and have obtained the corresponding detection performance. In IDS, the ANN is implemented by training neurons with the sequence of log audit files and sequence of commands [14]. ANN is used for classification of unauthorized users from authorized users. The process of intrusion detection with the help of Neural networks take place in two phases.  

Training: In the first phase we trained our network so that it can recognize normal and intruder behavior. Most of researcher has worked on the KDD99 dataset and it is publically available.  

Testing: The next phase after training is testing phase where our network model is tested on test dataset. This dataset is smaller than the training dataset to ensure that the network can detect intrusions it was trained to detect.

Literature review: The study of misuse detection based approach began with Anderson’s report in 1980, where he presented a report on automatic intrusion detection model. He was the first to propose the idea of using audit trails to track computer misuse and match patterns of use behavior [15]. In 1987, Dorothy Denning proposed the first intrusion detection model which became the baseline for most of the intrusion detection systems of today [16]. Fox et al presented their work in the field of intrusion detection to characterize normal and abnormal user using Self organizing map model [17]. Hua Tang et al proposed an approach to detect normal user from intruder using feed forward artificial neural network and support vector machines. They compared neural network model with support vector machine and shows that support vector machine have better results than neural networks [18]. Debar et al presented a new model of Intrusion detection which was based on recurrent neural network and has better results than neural network model [19]. Li used an Adaptive resonance theory neural network model for detecting normal behavior of legitimate user. All the data about a particular user access behavior is collected which is used in future to classify a user as normal or intruder [20]. Zhang Wei et al proposed a new model based upon back propagation neural network for anomaly detection and used supervise learning technique [21]. Barrus et al. designed a neural network for measuring and determining alert threshold values. And such alerts are interchanged between collections of autonomous agents. Hierarchical Intrusion Detection (HIDE) is another proposed model that employs neural networks for anomaly detection. HIDE makes use of statistical models and neural network classifiers to detect attacks [22]. Cannady came up with a new model based upon Multi Layer Perceptron (MLP) for identifying misuses in computer systems which has four fully connected layers. Training and testing data sets are generated using a network monitor tool [23]. South Korean researchers at Yonsei University have used evolutionary learning neural networks (ENN) to improve anomaly detection performance based on learning program’s behavior (Han at al, 2004). Their model made use of system-call audit data to build ENN normal behavior profiles. One neural network was used per program with 10 input nodes, 15 hidden and 2 output nodes (normal and attack). Anomaly detection performance reached almost 100% with very low false positive rates, whereas the structure and the weights of neural network were learned simultaneously. Nevertheless, the proposed model did not provide misuse detection, limiting success to anomaly detection [24]. Mukhopadhyay et al. proposed Back propagation neural network for intrusion detection. Their emphasis is on detection of new attacks and low failure rate and they use NSL KDD data set for experimental purpose. Table II shows different instances of data set present in training and testing data set of NSL KDD data set.

Figure 5: A general ANN model
networks have already been applied successfully to the various research domains including intrusion detection system. However current research work shows that there are still unsolved issues including false positives and negatives, detection of new attacks and human interaction dependence [25]. Also, most of the proposed models do not integrate together anomaly and misuse detection techniques. Snap-drift seems like a potential neural network algorithm candidate that might provide the final solution of interacting misuse and anomaly detection.

Table II. Instances of data set

<table>
<thead>
<tr>
<th>Class Type</th>
<th>Instances in Training Data set</th>
<th>Instances in Testing Data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>67343</td>
<td>9711</td>
</tr>
<tr>
<td>Dos</td>
<td>45927</td>
<td>7456</td>
</tr>
<tr>
<td>Probe</td>
<td>11656</td>
<td>2421</td>
</tr>
<tr>
<td>U2R</td>
<td>52</td>
<td>200</td>
</tr>
<tr>
<td>R2R</td>
<td>995</td>
<td>2756</td>
</tr>
</tbody>
</table>

**Fuzzy Logic:** The past decades have witnessed a rapid growth in the number and variety of applications of fuzzy logic. **Fuzzy Logic** introduced by Zadeh (1965) gives us a language, with syntax and local semantics, in which we can translate our qualitative knowledge about the problem to be solved. Fuzzy logic is an innovative technology which enhances conventional system design with engineering expertise. The use of fuzzy logic can help to circumvent the need for mathematical modeling. Anything that can be built using conventional design techniques can also be built with fuzzy logic, and vice-versa. The main issue with traditional fuzzy systems is that these are nonadaptive in nature. Therefore, building fuzzy systems with learning and adaptation capabilities has gained much interest recently [26].

By linguistic terms, intrusion detection features can be viewed easily and decision of normal and abnormal activity in the network are based on its fuzziness nature that can identify the degree of maliciousness of a node instead of yes or no conditions [27]. The fuzzy logic has been applied in IDS in the past for two major reasons, Firstly several quantitative parameters that are used in the context of intrusion detection, e.g., CPU usage time, connection interval, etc, can potentially be viewed as fuzzy variables. Secondly, the concept of security itself includes some degree of fuzziness. Classical methods that have been applied in the Intrusion detection system define an interval or a range of values to denote a normal user and intruder. Based on these approaches, any of the values that fall out side of the range are considered as anomalies regardless of their distance to the interval. Due to this abrupt separation between normal and anomaly false alarm rate significantly increases. Fuzzy logic helps to smooth this abrupt separation and provide flexibility in the IDS’s. The Fuzzy Logic uses the fuzzy variable along with the membership function to determine whether a particular rule is applicable to classify the condition as an anomaly or not. Fuzzy logic has proved to be a powerful tool for decision making to handle and manipulate Imprecise and noisy data [28]. Fuzzy based IDS’s is also able to model complex and non-linear problems and has natural language processing capability. The application of Fuzzy logic in IDS’s has the following representation.

**If condition then consequence**

Where

- **Condition** is a fuzzy variable.
- **Consequence** is the fuzzy set.

For better understanding consider a scenario, If number of requests with same destination address is **HIGH**, then IDS can consider it unusual. Now to determine how many requests will be considered in the category of HIGH, for this we have divided the number of requests into discrete sets known as fuzzy sets.

**Fuzzy Logic based Intrusion Detection System**

Fuzzy logic is used in intrusion detection since 90’s because it is able to deal with uncertainty and complexity which is derived from human reasoning. Fuzzy logic dealing with the vague and imprecise is appropriate for intrusion detection. By the help of fuzzy variables or
We have considered a fuzzy space of three sets LOW, MEDIUM and HIGH, then region X depicts fuzzy set LOW, Y depicts fuzzy set MEDIUM and Z depicts fuzzy set HIGH. The x axis shows the values in the fuzzy set and the y axis shows the membership function. The number of requests is the fuzzy variable which is also known as fuzzy linguistics whereas the LOW, HIGH and MIDDLE depict the values of the fuzzy variable. So if intrusion detection system receives requests in the interval 10 to 20 for membership function value 0.4 then it will fall in the category of LOW but for degree 0.6 it will be considered as HIGH.

So in the Intrusion Detection System using Fuzzy Logic, this can be written as,

\[
\text{IF Number of Requests = HIGH}
\text{THEN Abort Connection.}
\]

Number of Requests is the fuzzy variable and HIGH is the fuzzy set. So, depending upon the number of packets in the incoming traffic and the membership function, the value of HIGH is determined and accordingly the Intrusion Detection System will decide whether to abort the connection or not.

**Literate Review:** Many researchers have done a significant contribution in the field of Intrusion Detection System using Fuzzy Logic. Chavan et al. used Fuzzy Inference System combined with Artificial Neural Networks for real time traffic analysis [29]. Dickerson et al. developed the Fuzzy Intrusion Recognition Engine (FIRE) using fuzzy sets and fuzzy rules. They have used various parameters for intrusion detection example, CPU usage time, activity frequency, connection interval, etc., is fuzzy in nature [30]. Siraj et al. proposed a Fuzzy Cognitive Map (FCM) for capturing causal knowledge that decision engines use for alert assessment in computer network environments [31]. Cho used a fuzzy rule reasoning mechanism in order to detect an anomaly. The input of the fuzzy reasoning mechanism are Hidden Markov Model (HMM) evaluation values (from different HMM models). A centroid defuzzification technique is applied for determining the final classification (abnormal or normal) [32]. Tajbakhsh et al. proposed a system to classify normal and anomalous attacks on the basis of compatibility threshold. Here the researchers used association based classification to classify the network traffic data [33]. Barbara et al. proposed Audit Data Analysis and Mining (ADAM) which was a real time anomaly detection system. Here, suspicious events were classified as false alarms or real attacks by a module using association rules along with data mining techniques and classification [34]. Botha et al. report a work to detect the intrusion using the user behavior and the fuzzy logic methodology. In this paper, intrusion detection algorithms are similar to the two earlier approaches introduced in previous papers. The overall view of the authors is to consider six different generic phases for an intrusion into a network. The goal of this system is to track and monitor the current state of the user’s behavioral profile considering these categories [35].

Although Fuzzy Logic based system is very efficient and finds a wide range of applications in various fields including Intrusion detection systems, but it also has some shortcomings enlisted as follow

- Inability of self learning
- Adaption or parallel computation
- Cannot support optimization
- Answer obtained once cannot get better with time.

**Genetic Algorithm:** “A Genetic Algorithm (GA) is a programming technique that mimics biological evolution as a problem-solving strategy. It is based on Darwinian’s principle of evolution and survival of fittest to optimize a population of candidate solutions towards a predefined fitness. Genetic Algorithms are biologically inspired search heuristics that employs evolutionary algorithm techniques like crossover, inheritance, mutation, selection etc. So, genetic algorithms are capable of deriving classification rule and selecting optimal parameters for detection process. GA uses an evolution and natural selection process to search for good solutions to a given problem. It is a population based technique that involves the application of evolutionary operators such as mutation, crossover and selection to a population of candidate solutions to a problem. The goal of GA is to find the optimal solution or a solution that is close to the optimal one.

![Figure 6: Fuzzy space used in Intrusion Detection](image-url)
selection that uses a chromosome-like data structure and evolve the chromosomes using selection, recombination, and mutation operators. The process usually begins with randomly generated population of chromosomes, which represent all possible solution of a problem that are considered candidate solutions. Different positions of each chromosome are encoded as bits, characters or numbers. These positions could be referred to as genes [36]. The literature demonstrates that the Genetic Algorithms provide better and faster classification than any neural network architectures, and also takes less time for training and gives detection rate. Genetic algorithm based intrusion detection system is used to detect intrusion based on past behavior. A profile is created for the normal behavior based on that genetic algorithm learns and takes the decision for the unseen patterns. Genetic algorithms also used to develop rules for network intrusion detection.

**Parameters used in genetic algorithm**

- **Fitness Function**: The fitness function evaluates the quality of a particular solution. The fitness function is used to select the best solution among all the solutions in the population. The fitness function should be an optimized value.
- **Selection**: Selection is the process of choosing solution with better fitness function than their counterparts. In the selection phase the solutions having better fitness function over other solutions are selected and the rest are discarded.
- **Crossover**: Crossover is the phase in which two solutions exchange one of their characteristics with the other in the pair at a randomly selected crossover point, where the crossover probability is between 0.6 and 0.9. The solutions selected for crossover operation should be different.
- **Mutation**: Mutation is a process by which some random bits in a solution are changed. This is done mainly to maintain the genetic diversity of the solutions [37].

When genetic algorithm is used for problem solving, three factors will have impact on the effectiveness of the algorithm, they are

a. The selection of fitness function
b. The representation of individuals and
c. The values of the genetic parameters

**Major Steps in Genetic Algorithm**

**Algorithm**: Rule set generation using genetic algorithm.

**Input**: Network audit data, number of generations, and population size.

**Output**: A set of classification rules

1. Initialize the population
2. Check the fitness function
3. Select only those rules that that meets the fitness criteria.
4. Perform crossover for reproduction of new rule by exchanging some bits
5. Perform mutation by flipping some bits
6. Again go to line 2, until the specified numbers of rules are not generated.

**Data representation in genetic algorithm**: Genes should be represented in some format using different data types such as byte, integer and float. Also they may have different data ranges and other features, knowing that the genes are generated randomly, in each population generating iteration. The conditions to detect the intrusion is generally the current network traffic or connection details like source IP address, destination IP address, port numbers (like TCP, UDP), duration of the connection, protocols used. For genetic algorithm to work in IDS each of above mentioned condition is converted into chromosome. Each chromosome is evaluated by a fitness function to determine the solution's quality; better-fit solutions survive and produce offspring, while less-fit solutions are culled from the population.
algorithms can be used to evolve rules for the network traffic; these rules are usually in the following form:

“If {condition} then {act}”

It basically contains if-then clause, a condition and an act. The conditions usually matches the current network behavior with the one stored in the in the IDS such as comparing an intruder source IP address and port number with one already stored in the system. The act could be an alarm indicating that the intruders IP and Port numbers are related to an attacker who is previously known in the system. Example

if (duration = “0:0:1” and protocol = “finger” and source_port = 18989 and source_ip = “99.19.99.19” and destination_ip = “192.168.254.10” ) then (attack name = “Neptune”). Above rule specifies that if a network packet is originated from IP address 99.19.99.19 and port number 18989 and send to IP address 192.168.254.10 at port number 79 using finger protocol for duration of connection 1 second then most likely it is Neptune attack which eventually make destination host out of service.

**Literature review:** Several case studies and experiments were applied regarding the use of genetic algorithm in intrusion detection system. Several Genetic Algorithms and Genetic Programming have been used for detecting intrusion detection of different kinds in different scenarios. Chittur at al. applied detection of computer intrusions and malicious computer behavior and analyzed the effectiveness of a Genetic Algorithm and compared their results with previous intrusion detection techniques [38]. Li described a method using GA to detect anomalous network intrusion. His approach includes both Quantitative and categorical features of network data for deriving classification rules. He addressed the Factors affecting to Genetic Algorithm are in detail. The inclusion of quantitative feature can increase detection rate but experimental results are not available [39]. Xia et al. detected anomalous network behaviors based on information theory by using GA. Some network features can be identified with network attacks based on mutual information between network features and type of intrusions and then using these features a linear structure rule and also a GA is derived. The advantage of the approach of using mutual information and resulting linear rule seems very effective because of the reduced complexity and higher detection rate. The approach has disadvantage that it considered only the discrete features [40]. Abdullah showed a GA based performance evaluation algorithm to network intrusion detection. The traffic data was filtered by information theory in his approach [41]. Lu and Traore used support-confidence framework as the fitness function and accurately classified several network intrusions. Disadvantage of their approach is that, use of genetic programming made the implementation procedure very difficult and also for training procedure more data and time is required [42]. Gong et al. presented an implementation of GA based approach to Network Intrusion Detection using GA and showed software implementation. In this approach he derived a set of classification rules using a support-confidence framework to judge fitness function [43].
Table III: Comparison of Existing Studies on GA Based IDS

<table>
<thead>
<tr>
<th>Authors</th>
<th>Detection approach</th>
<th>Fitness function (F)</th>
<th>Explanation of Fitness function used</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.A. Ojugo, A.O. Eboka, O.E. Okonta, R.E Yoro (Mrs), F.O. Aghware</td>
<td>Misuse analysis</td>
<td>Support and confidence model F=W1<em>support+W2</em>confidence</td>
<td>If we have the rule: If A then B, support =</td>
<td>A and B</td>
</tr>
<tr>
<td>B. Uppalaiah K. Anand, B.Narsimha, S.Swaraj, T.Bharat</td>
<td>Misuse analysis</td>
<td>Fitness = f(x) / f (sum)</td>
<td>Where f(x) is the fitness of entity x and f is the total fitness of all entities</td>
<td>Uses only 3 network features; 83.65% of avg. success rate; process is faster, can be applied for high speed networks</td>
</tr>
<tr>
<td>Bharat S. Dhak, Shrikant Lade</td>
<td>Misuse analysis</td>
<td>F= weight*packet_size</td>
<td>Where the packet_size is the actual packet data size prescribed by the incoming packet data stream and weight is the Vector which is applied to each chromosome</td>
<td>Scope of experiment is focused to generate a list of vulnerable IP addresses; gained 96% of accuracy.</td>
</tr>
<tr>
<td>Firas Alabsi, Reyadh Naoum</td>
<td>Misuse analysis</td>
<td>Reward Penalty model based F=2+(AB-A/AB+A)+(AB/X)-(A/Y)</td>
<td>Consider a rule: If A then B, ((AB-A)/(AB+A))= strength of a record; AB/X= ratio of the strength of record to the strength of the strongest record; A/Y=ratio of the weakness of a record to the weakness of the weakest record;</td>
<td>Uses 5-network features; Fitness function gives reward to good chromosomes and applies penalty on the bad chromosomes; comparison between the newly proposed and other existing fitness functions is presented.</td>
</tr>
<tr>
<td>Wei Li</td>
<td>Anomaly detection</td>
<td>Weighted sum model based F=1-penalty</td>
<td>Fitness function is determined by calculating the general outcome, absolute difference and penalty values.</td>
<td>Considers both temporal and spatial features of a network connection to detect an attack; no experimental results</td>
</tr>
</tbody>
</table>
So from the literature review it is observed that GA provides following major advantages

- GA can be easily interfaced to obtainable simulations and models
- GA is easy to hybridize and easy to understand
- GA uses little problem specific code
- GA is modular, separate from application
- GA is capable to obtain answers always and gets better with time
- GA is inherently parallel and easily distributed.
- GA can be used to find optimized values from large search.

**Limitation of GA:** The GA is efficient but in practice they have certain limitations.
- It is not capable to store domain knowledge.
- It is not always easy to find a fitness function.
- Representing a problem space in genetic algorithms is very complex.
- It is a tough task to choose the optimal parameters for a genetic algorithm.
- Genetic algorithms need a large number of fitness function evaluations.
- It is not easy to configure a genetic algorithm based system.

### Intrusion Detection using Hybridization of Soft Computing Techniques

From last two decade soft computing techniques had been applied individually in intrusion detection system, but each of these techniques has its own advantages and disadvantages.

Many researchers have suggested that the monitoring capability of current IDS can be improved by taking a hybrid approach that consists of both anomaly as well as signature detection techniques (Lunt et al. 1992; Anderson et al. 1995; Porras and Neumann 1997; Hwang et al. 2007; Fortuna et al. 2007). So researchers come up with an idea of hybridization of soft computing techniques like Neuro-Fuzzy, Fuzzy-Genetic, Neuro-Genetic and Neuro-Fuzzy-Genetic. Neuro-Fuzzy approach is useful and very much effective in intrusion detection where first neural network is trained on fuzzy inputs and then it provides explanation and reasoning. Thus hybrid methodology is helpful in overcoming the black box nature of artificial neural network which is a major disadvantage of neural network. Researchers have developed a cooperative Neuro-fuzzy approach where the system first uses neural network to decide and enhance parameters of fuzzy logic system, once parameters and rules are defined, the fuzzy logic system takes the charge and solve the problem more effectively. In many applications multilayer neural networks are used to extract fuzzy rules.

### Table IV: Comparisons of ANN and Fuzzy inference system

<table>
<thead>
<tr>
<th>Artificial Neural Network</th>
<th>Fuzzy Inference System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior rule-based knowledge cannot be used</td>
<td>Prior rule-base can be incorporated</td>
</tr>
<tr>
<td>Learning from scratch</td>
<td>Cannot learn (use linguistic knowledge)</td>
</tr>
<tr>
<td>Black box</td>
<td>Interpretable (if-then rules)</td>
</tr>
<tr>
<td>Complicated learning algorithms</td>
<td>Simple interpretation and implementation</td>
</tr>
</tbody>
</table>

Fuzzy and genetic systems are hybridized in the following ways

- Genetic algorithms controlled by fuzzy logic, where genetic operators and functions use fuzzy linguistic terms
- Fuzzy evolutionary systems, where fuzzy rule sets are evolved

The fitness function of genetic algorithms may employ fuzzy variables to determine suitable candidates within the population and effective size of populations [44]

From results of these hybrid techniques it has been cleared that hybrid approaches are superior more accurate and more capable to detect intrusion than separate used of approaches as ANN, Fuzzy logic and Genetic. Ensemble and hybrid techniques are becoming increasingly popular now days. Accuracy and false positive rate which are important parameter in any IDS has been improved by applying hybrid approaches that consists of both anomaly as well as misuse detection techniques. A hybrid approach involves integration of different learning or decision-
making models. Each learning model works in a different manner and exploits different set of features. Integrating different learning models gives better performance than the individual learning or decision-making models by reducing their individual limitations and exploiting their different mechanisms. The benefits of hybrids approaches include robustness, improved performance and increased problem-solving capabilities. Finally, fully hybrid models can provide a full range of capabilities such as adaptation, generalization, noise tolerance and justification. One of the problem with hybrid systems is increased complexity of the inter-module interactions and specifying, designing, and building fully integrated models is complex.

Table V: Strengths of soft computing components

<table>
<thead>
<tr>
<th>Field</th>
<th>Strengths offered</th>
</tr>
</thead>
</table>
| ANN   | • Learning and implicit knowledge representation.  
       • No prior knowledge of the data generating process is needed for implementing NN  
       • Problem of model misspecification does not occur |
| FL    | • Simplicity, flexibility and approximate reasoning  
       • Fuzzy logic can handle problems with imprecise and incomplete data.  
       • It can model nonlinear functions of arbitrary complexity. |
| GA    | • Natural evaluation and optimization  
       • GA is easy to hybridize and easy to understand  
       • GA is capable to obtain answers always and gets better with time. |

(Peddabachigari et al. 2007) Common objective of ensemble methods is to construct some combination of some models, instead of using a single model to improve the results [45]. Mukkamala et al. (2004) have proved that by using of ensemble classifiers best accuracy for each category of attack patterns can be achieved [46]. Ensemble approach of different classifier tries to improve the predictive performance of learnt model. Chebrolu et al. (2005) utilized CART-BN approach for intrusion detection (Chen et al. 2005). CART performed best for Normal, Probe and U2R and the ensemble approach worked best for R2L and DoS [47]. Abadeh et al (2007) described the fuzzy genetic-based learning algorithm and discussed its usage to detect intrusions. They presented some results and reported the performance of generated fuzzy rules in detecting intrusion in a computer network [48]. Koutsouros et al (2007) presented a neural network classifier ensemble system using a combination of neural networks which is capable of detecting network attacks on web servers. The system can identify unseen attacks and categorize them. The performance of the neural network in detecting attacks from audit dataset is fair with success rates of more than 78% in detecting novel attacks and suffers from high false alarm rates [49]. Zainal et al. (2009) proposed the ensemble of Linear Genetic Programming (LGP), Adaptive Neural Fuzzy Inference System (ANFIS) and Random Forest (RF) for ID. They have empirically proved that by
assigning proper weights to classifiers in ensemble approach improves the detection accuracy of all classes of network traffic than individual classifier [50]. Menahem et al. (2009) have utilized multiple different classifiers and have tried to exploit their strengths. Authors used KDD data set and experimental study shows ensembling of classifiers gives different accuracy rate on different classes of attack. The success of an ensemble method depends upon many factors, including the training sample size; the choice of a base classifier; the exact way in which the training set is modified; the choice of the combination method; and, finally, on the data distribution and the potential ability of the chosen base classifier to solve the problem [51]. Wang et al (2010) proposed IDS based on ANN and fuzzy clustering, which helps to achieve higher detection rate, less false positive rate and stronger-showed stability [52]. Hwang et al. (2007) has proposed a 3 tier hybrid approach to detect intrusions. First tier of system is signature based approach to filter the known attacks using black list concept. Second tier of system is anomaly detector that uses the white list concept to distinguish the normal and attacks traffic that has by passed first tier. Third tier component of system uses the SVM to classify the unknown attack traffic into five classes i.e. normal, probing, DoS, U2R and R2L. KDD dataset was used to train and test the system [53]. The results are summarized in following

**Table VI: Summary of results (Hwang et al. 2007)**

<table>
<thead>
<tr>
<th>Class detection/Accuracy</th>
<th>Old attacks %</th>
<th>New attacks %</th>
<th>Total detection %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probing</td>
<td>99.92</td>
<td>98.16</td>
<td>99.14</td>
</tr>
<tr>
<td>DoS</td>
<td>99.99</td>
<td>18.03</td>
<td>97.65</td>
</tr>
<tr>
<td>U2R</td>
<td>20.57</td>
<td>87.83</td>
<td>76.32</td>
</tr>
<tr>
<td>R2L</td>
<td>79.84</td>
<td>26.94</td>
<td>46.53</td>
</tr>
</tbody>
</table>

**Table VII: Summary of results (Menahem et al. 2009)**

<table>
<thead>
<tr>
<th>Class Name</th>
<th>Correctly Classified</th>
<th>TPR</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>97.67</td>
<td>0.97</td>
<td>0.93</td>
<td>0.925</td>
</tr>
<tr>
<td>Probing</td>
<td>92.83</td>
<td>0.92</td>
<td>0.88</td>
<td>0.82</td>
</tr>
<tr>
<td>DoS</td>
<td>74.65</td>
<td>0.74</td>
<td>0.68</td>
<td>0.92</td>
</tr>
<tr>
<td>U2R</td>
<td>86.45</td>
<td>0.86</td>
<td>0.76</td>
<td>0.88</td>
</tr>
<tr>
<td>R2L</td>
<td>98.86</td>
<td>0.98</td>
<td>0.92</td>
<td>0.82</td>
</tr>
</tbody>
</table>

TPR here is true positive rate.

Wang Yunwu (2009) analysed the current situation of IDS by using genetic based fuzzy system algorithm. He presented a genetic fuzzy expert system (GFES) in which he applied association rules together with fuzzy logic to classify the data. The advantage of this is it needs less fuzzy rules to achieve a certain high rate of recognition and classification. Fuzzy rules do not need any luminous knowledge for generating rules [54]. Fan Li (2010) proposed IDS based on hybridization of artificial neural networks and genetic algorithm. Input features, network structure and connection weights are evolves using genetic algorithm. This is helpful for identification of complex anomalous behaviors [55]. Shaokun Liu et al. (2012) presented an improved GA to solve the optimization problem and also introduced the greedy algorithm for better improvement. The advantage of this method is the performance of improved GA better than simple GA [56]. Mostaque Md.et al. (2013) has presented methodologies and good fuzzy classifier using GA, he also presented challenges in IDS. Author proposed the new definition of fuzzy set where he described the fuzzy membership value and fuzzy membership function for the complement of a fuzzy set are two different concepts because the surface value is not always counted from ground level [57]. The advantage of this technique helps in reducing the false alarm rate in IDS. Jongsuebsuk (2013) proposed a new framework for unknown or
new network attack types with a help of fuzzy genetic algorithm technique. The Fuzzy Genetic Algorithm is rule-based which does not require high computation time. The results show that the system has high detection rate and average detection rate is approximately 97%.

**CONCLUSION:** An attempt has been made in our research to explore the essential of Intrusion Detection System using soft computing. We have seen the advantages and disadvantages of each of soft computing techniques in IDS when applied individually, which leads research towards ensemble approach. An ensemble helps to indirectly combine the synergistic and complementary features of the different learning paradigms without any complex hybridization. The ensemble approach successfully classifies all the five classes of NSL KDD data set. It has been found that if proper soft computing paradigms are chosen, their ensemble might help in gaining 100% classification accuracies. Since all the considered performance measures could be optimized such systems could be helpful in several real world applications.

**REFERENCES:**


21) Zhang Wei, & Wang Hao-yu (2010), Intrusive Detection Systems Design based on BP Neural Network, IEEE.


41) B. Abdullah, I. Abd-alghafar, Gouda I. Salama, A. Abd-alhafez, “Performance Evaluation of a Genetic Algorithm Based