CYBER THREAT DETECTIONTECHNIQUES

Sanjana RS [1], Sandeep Esap [2], Rajeshwari K [3]
[1,2]Informaation Science and Engineering BMS COLLEGE OF ENGINEERING Bangalore, India
[3]Assistant Professor,ISE BMS COLLEGE OF ENGINEERING Bangalore, India
Email: 1bm16is78@bmsce.ac.in, 1bm16is76@bmsce.ac.in, rajeshwarik.ise@bmsce.ac.in

Abstract— As we are entering the phase where the world runs on internet and most of our important transactions to work is carried out online there is always a threat of being into fraud caused by online methods. Many forms of online fraud have been carried out like phishing, spear phishing, malware and Trojans. Cyber threat detection is a technique of learning about the threats and then making itself ready to fight it back for the next time a similar attack is encountered. This process uses software techniques of Data Mining, Machine Learning and Artificial Intelligence methods and techniques and algorithms to collect previous data and learn from it and make itself ready for any future attacks that might happen. The following paper talks about the algorithms trained by the DARPA 1998 dataset. Data Mining uses algorithms like Association rule, clustering, decision trees, neural network and statistical techniques and Machine learning uses Bayesian network, Decision tree, genetic algorithms, Hidden Markov models, Inductive learning and reinforcement learning. Artificial Intelligence techniques include intelligence agents, learning and constraint finding.
Keywords→Intrusion detection, Decision trees, Artificial neural networks, Clustering, Association rules, Reinforcement learning.

I. INTRODUCTION

Cyber security are methods, practices and algorithms adopted for protection of important information in computer systems, devices, electronic systems and networks from being attacked from fraud methods. Cyber threat detection are means of finding out the ways in which a hacker can intrude in a system or network and their behavior adopted to do so and find means and algorithms to prevent it from such further attacks.In this paper we use machine learning, data mining and artificial intelligence algorithms to train the previous network data and keep it ready for future attack detection.We can relate Machine Learning and data mining as they share similar set of methods and overlap often .The analysis of knowledge discovery of the database and finding unknown values and properties of the data given which was not previously known while researching is the main goal of Data Mining, prediction of properties learned from the already existing training data set is the way Machine learning works. Data Mining having a different set of goals as compared to machine learning it still uses few of the machine learning methods and similarly Machine Learning also uses unsupervised learning or pre-processors from the methods used by Data Mining for increasing the learning capability and their accuracy. They have a lot of assumptions regarding the confusion among ML and DM, these do contain different journals and conferences except for PKDD and EMCL: The discovery of the previous unknown information and data is the main aim of Data mining and the ability to learn and reproduce the already existing knowledge which different data sets after being trained by a training dataset is the goal of Machine learning. Having a shortage or unavailability of the training data supervised methods will not be effective in a KDD task, many of the different supervised methods can easily outperform most of the unsupervised methods when being evaluated in respect to known knowledge. Artificial Intelligence which is the super set of Machine learning and Data Mining also uses learning of data from traffic collected and respond to future such events and are capable of detecting for new set of intruding traffic as well and here we have added up AI tools also like expert systems which follow the similar procedure of training itself or learning from previously collected data. This is more visualized like the neurons and in brains and the connections across them to improve our thinking ability and perform actions according to the environment around us.

II. LITERATURE SURVEY

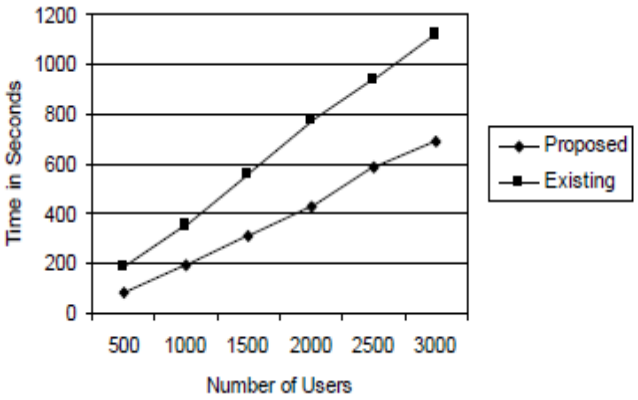
Data mining, ML and AI has been used extensively in the field of cyber intrusion detection and is very effective as it learns from previously collected traffic data and can train itself for future threats as is even capable of predicting other possible ways of intrusion from the previously collected data. Hanen Brahmi, Imen Brahmi and Sadok Ben Yahia [1] propose a way to detect a threat to a system using a multidimensional association rule mining which follows a way of detecting correlation relationships among data. The paper firstly looks at the data and finds out all the rules of the type X → Y where X is the antecedent and Y is the consequent of the rule, which is good enough for the specified minimum support "minSup" and the minimum confidence "minConf". It considers two multidimensional association rules D1 and D2 where D1 has {Source port = 22, Destination port= 193.62.10.10 and service type= telnet and duration =long} then the attack type could be classified as {Smurf } and D2 has {Source port = 22 and service type = telnet} then the attack would be {Smurf} .As we see that D1 and D2 share similar features of the source port as 22 and the service type as telnet we can say that D2 is redundant with respect to D2. This helps in any future detection of similar threats. This method is proven to be quite effective in accuracy and less false positive rates.

S. Sathya Bama, M.S. Irfan Ahmed and A. Saravanan [2] have proposed the agglomerative clustering algorithm for cyber threat or intrusion detection. Where a data when received along with its subset of features is classified to the most approximate set of clusters by calculating the K-nearest neighbour. Any of the data which falls outside the clusters and have a threshold below the specified amount can trigger an alarm saying it is an intrusion. There are many clusters and the data in one cluster are similar to each other

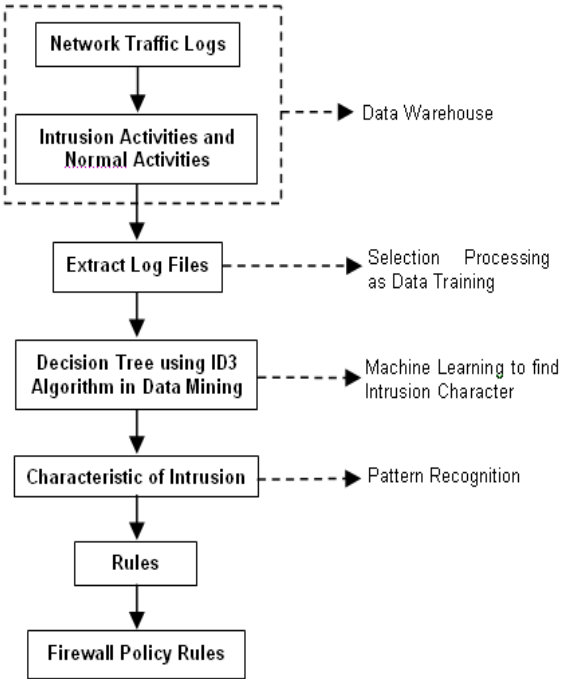
having similar features is completely dis-similar to the features of the other cluster. The procedure has to look at common features among the data with the similar order as given in the set of clusters and continue to do the same for all clusters till it fits into one. To calculate the similarity, a formula sim (Sq1, Sq2) =|E3| ∩ |E4| / |E1|+|E2|/2 and the criterion function

$$\text{Maximize } C_f = \sum_{r=1}^{k} \frac{1}{n_r} \sum_{i,j=c} \text{sim}(i, j)$$

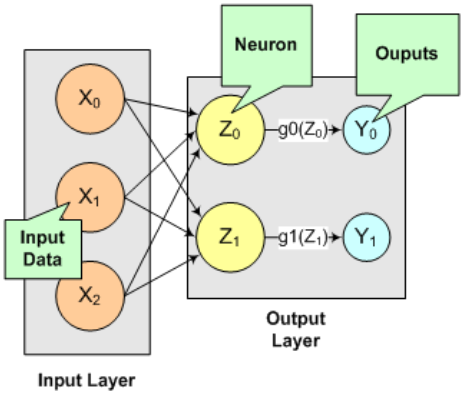
The experimental results with a sample dataset give the graph



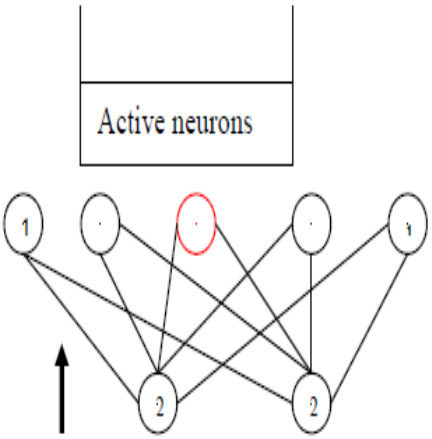
Phuong Do, Ho-Seok Kang and Sung-Ryul Kim [3] proposes ID3 decision tree algorithm to classify the data from normal and intrusion considering that the behaviour of the intruder is mostly different from that of a legitimate user. The decision tree is used to classify the rules into subsets as comparing small set of attributes with a large set of rules is a complex procedure. It collects the normal behaviour of the legitimate user and classifies it accordingly and any set of user actions that fall out of the legitimate user path is marked as an intruder.



E. Kesavulu Reddy [4] talks about the Neural network based intrusion detection systems. These usually have a layers and data travels from one layer to another working with each other just like the brain process. It talks about multilayer feed forward network which has 3 layers – input, middle or hidden and the output training itself to produce the outputs corresponding to the input patterns and gives output to unknown data corresponding to the previously taught input. It uses backpropagation to train itself for unknown input data.



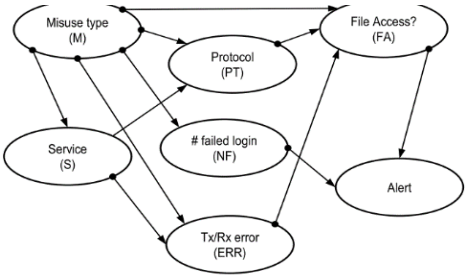
The other method is Kohonen's self-organizing Map used for unsupervised learning and clustering the data where the connections between layers have weights assigned and updated according to the kohonen's rule only to topological neighbours and active output neurons and the best node is well trained to give the required output for a given input.



1-- Output Layer

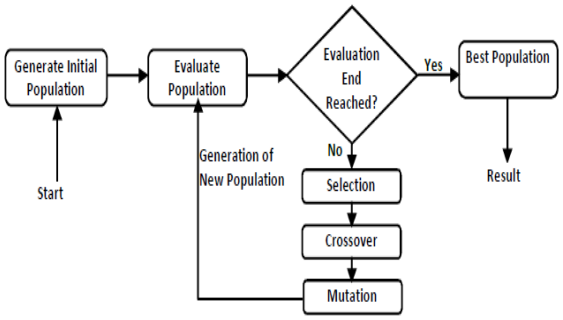
2--- Input Layer

Farah Jemili, Montaceur Zaghdoud and Mohamed Ben Ahmed [5] proposed a Bayesian network for intrusion detection system considering an acyclic connected graph where the nodes are child nodes connected to parent nodes and the state of the child node is dependent on the parent node. Each of the nodes in the graph have a conditional probability of the variable given by the formula P(B/A) = (P(A/B) P(B))/P(A). The state of the graph at a given time shows the activity of a user and can be differentiated from that of an intruder as they will have a different set graph states when they use the system. Here is an attack signature detection table using Bayesian network.



File Access state input variables and values	P(FA = True)	P(FA = False)
M=R2H, PT=NSF, ERR=0	0.95	0.05
M=R2H, PT=FTP, ERR=0	0.99	0.01
M=Probe, PT=none, ERR=50%	0.80	0.20
M=Probe, PT=PING, ERR=0	0.50	0.50
M=DoS, PT=POP, ERR=100%	0.80	0.20
M= DoS, PT=HTTP, ERR=50%	0.90	0.10

Parry Gowher Majeed and Santosh Kumar [6] In their paper talk about the usage of genetic algorithms for intrusion detection. It follows the process of fitness function which is used to find the best solution of the solutions for a particular intrusion. It is done by selection, crossover and mutation which is to find the most optimal solution and then cross it over with another optimal solutions to exchange the most required characteristics until the best possible solution can be determined for the set of problems detected and mutation is to make few changes to the solution to make it best fit the problem detected.



Joshi et. al. [7]used a hidden markov model for an intrusion detection system. It determines the topology of the model by using transition probabilities for interconnection between the states. It is useful in determining un-observed parameters from observed ones which make the model detect the unknownintrusion. Using six observation symbol per state five definite states are used here. Transitioning of states to one another is done with the kind of interconnection between the states. Baum-Welch method can be utilised for finding out the HMM parameters. This method used KDD 1999 dataset for experiments. Five features out of forty-one were used for analysis which resulted in 21% false positive rate and an 79% of positive detection rate. An assumption that using more features could have improved the results was claimed.

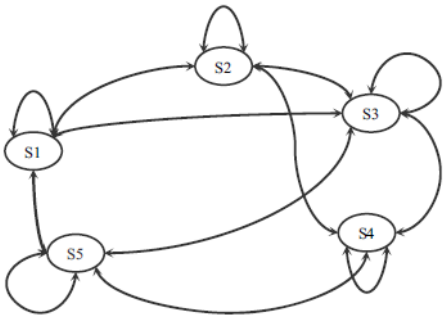
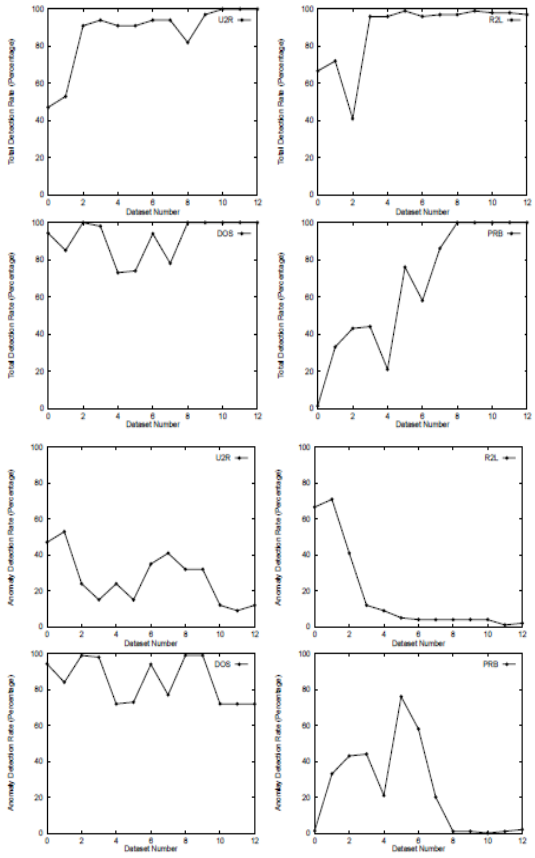
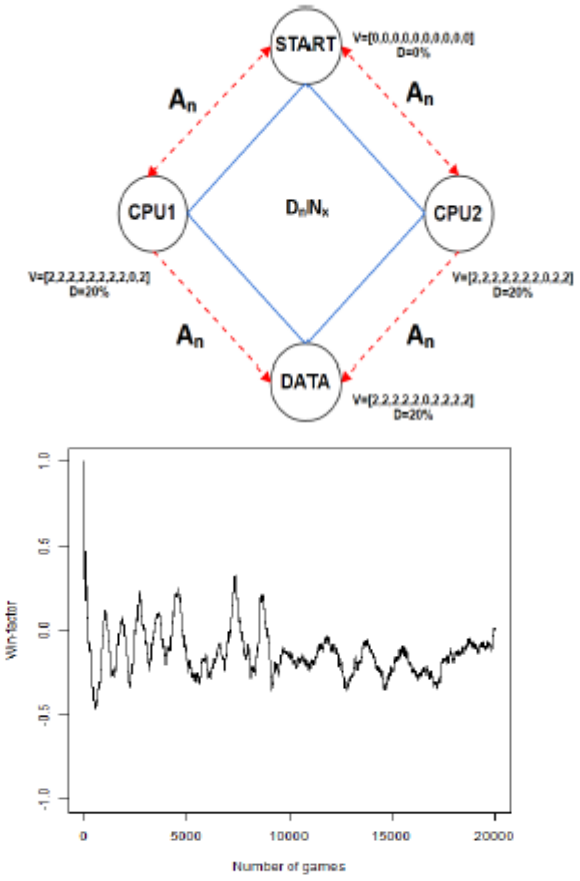


Figure 1. Finite State Automata of the State Transition from one state to another state for 5 features of KDD Cup 1999 data set

Fan et. al. [8]developed an artificial anomaly generator to generate random events and anomalous traffic. To develop this random anomaly usage of filtered artificial anomalies and distribution-based anomaly containing two prime approaches are used. DAPRA 1998 dataset was used to fuse the randomly generated data. This dataset was then used by him to understand and study the performance of inductive learning model developed. The experimental study resulted in 2% of Low far and 94% of successful detection rate which then helped to understand and develop a correct method of making a dataset that could be used for intrusion detection and put forth the application of inductive learning model and finally a graphical representation of the experimental results of total detection and percentage of known intrusion and true anomalies detected as anomalies.

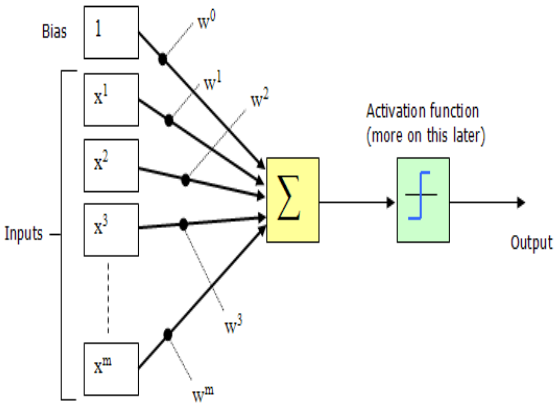


Richard Elderman [9] in his paper about reinforcement learning which uses the technique of self-learning by producing sample variables and performing computations on them to learn and train itself from a large range of data, It uses a markov game played between two people the attacker and the defender and the game is visualised as 3 layers of nodes where the start and the end are the attacker and defender and the middle ones are nodes through which they get to the other end and are connected by arrows. The defender has to goal of not allowing the attacker to get to the node where the defender resides as it is the system with sensitive information and the attackers goal is to attack the system where the defender is, and the game is won by the person who succeeds his goal. The defender has access to all nodes of the system where as the attacker doesn't, each of them proceed with the set of rules assigned to them to play the game and they have an attack and defend value which is incremented on each node when the person is successful in acquiring that node in case of the attacker and then proceeds to the next one to do the same and goes on until he reaches the destination and the defender increases his value in the node when he is successful in stopping the attacker from proceeding forward and prevent the system from being intruded. This helps the system to learn from its own mistakes and get ready for any future attack that could possibly have this set of procedures used by the intruder. This is sample representation of the game and experimental results as done in the paper for over 500 games.



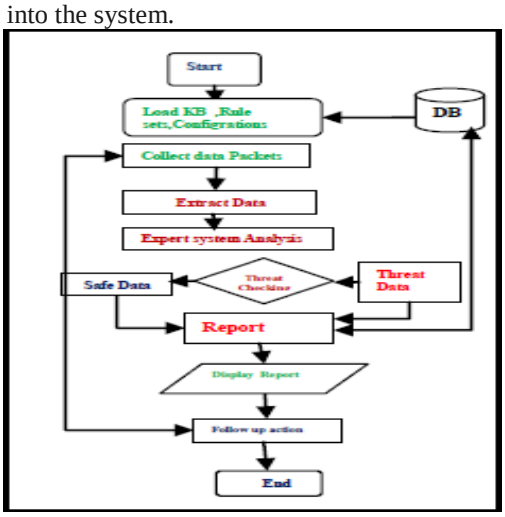
Xian Du [10] in his paper proposed the statistical methods of detecting intrusion detection which is quite simple as its main goal is to note down all the process of the legitimate user and the traffic generated over a long period by him initially and all the traffic data that falls outside of the legitimate user traffic is known as the traffic generated by the intruder and this approach has ambiguities as there can be exceptions in the users behaviour at times and this increases the false positive rates of the system and the longer the system is active and continues to work in detecting and learning the more accurate it becomes over time and so we say that though Data mining techniques are there for along period of time working in detecting threats in a system it hasn't been able to be completely successful and needs more improvements in the same field.

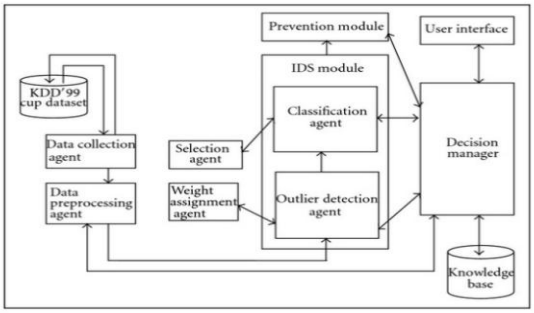
Amit Rajbanshi, Shuvam Bhimrajka, C. K. Raina[11] In the year 1958 Frank Rosenblatt invented perceptron which led to visual nets which are now used in Artificial Intelligence. The visual nets have neural nets in them that is usually looked at as a nerve cell. Perceptron has been trained to solve many problems and it can be applied as a part of visual nets to detect intrusion. Neural nets have embodied many of such artificial neurons which helps extensively in parallel learning from a provided dataset and decision making. This feature makes the work much faster than any of Machine Learning methods. They provide other benefits like pattern recognition, classification and selective response to attack types. They are suitable to be used in hardware and software systems and have a future scope of being used in various attack types like in DOS detection, spam detection, laptop worm and malware classification and in rhetorical investigations. They are having further enhancements in their features and usage like malware classification and in rhetorical investigations in the third generation.



Enn Tyugu[12] Expert Systems are on of the most used AI tools for intrusion detection. This system is used to search or locate answers to queries in an application domain conferred either by a user or by software. It is widely used in identification, finances and electronic network and for advance problems in this usage of diagnostic systems to a huge hybrid system is used. It includes a mental object with training data and a number of specific application domains square measure along with a part of system which is good at logical thinking for the given data. The systems having these two features is named as competent system shell which is filled with data beforehand with the help of an software system that could support it and is fed with data at intervals

to the mental object and then it can be extended with a few desirable programs for user interactions and systems that can be used in hybrid competent systems. The procedure of developing a competent associate system involves adaptation and selection of an associate competent system and then filling it with required data. It has additional advantages of simulation and calculation and the square measure huge amounts of data illustration forms in professional systems and the most commonly used is the rule-based illustrations. This system is mainly dependent on the quality of data given to it and the intervals in which they are fed into the system.





Guy G. Helmer, Johnny S. K. Wong, Vasant Honavar, and Les Miller[13] Intelligent agents in AI are those system which poses some extent of self-intelligence like pro activeness, reactivity, planning ability, quality and reflex ability along with understanding capacity of associate agent communication. The usage of agent communication language and objects that are minimum of proactive are said to be thought of in a software package agent by the software package engineering community. They are a combination of many monitors and network systems along with its network data recorded that detect the intrusion. They are majorly used in network systems and work by creating their own set of databases from the information collected from various sources regarding call traces and set of normal and abnormal behaviour. This paper even talks about the JAM that is (java agents for meta-learning) that can learn the data given and even exchange it with other systems to help them learn about the data. They have separate agents monitoring different parts like host, mobile agents and use data mining procedures for heterogeneous data. It identifies sources and gives an documentation for the systems administrators.

Chimphlee et al[14] and Dickerson et al[15] Fuzzy Logic is where instead of an accurate or an exact value the method of approximating it to the desired value by using a range of classifications are used which are similar to the ways a human brain thinks. The variables usually take two fixed values as true or false in traditional binary sets. A truth value ranging from 0 to 1 will be taken in case of fuzzy logicwhich helps in better classification of anything due to its high level of compactness, accuracy, flexibility and freedom in case of representing the real world situations. Most of the commonly used Boolean algebra properties can be used in fuzzy logic as well and the probability's degree of belief in a Boolean variable becomes a fuzzy variable's degree of truth. As humans are capable of giving different kinds of answer to a particular question like for its partially or completely cold for the whether temperature fuzzy logic is defined to impersonate such exact nature of the humans. Chimphlee et al using fuzzy c-means and rough set theory proposed an IDS which gave an output of around 93.45% OA using and KDD'99 dataset and an better performance was noticed when the count of the features was reduced. Dickerson et al for detecting the intrusion made a FIRE IDS that follows simple data mining techniques to learn the network data and expose the metrics. This is then utilized to detect the intrusion based on the learnt features.

Kim et al[16] and Zhang et al[17] Random Forests are a collection of trees giving larger range of classification fir the supervised learning algorithm based on the set of tree predictors. The selection from the training data set is used to decide where to grow each tree. Among a set of given classes, the selection of which one is said by the tree when it chooses. To make a set of weak learners be together to become a strong one is done by the forest that chooses the classification having the most votes over all the trees in the forest. Kim et al proposed that this method of random forests has been able to give out a stable result for the features and shows a high DR. The performance of the Random forests based on the IDS turns out to be comparable to the SVM. Zhang et al experiments showed a result of 92.58% of DR on the dataset of KDD'99 and on a balanced dataset a percentage of 99.86%, to increase the minority intrusions DR down sampling of the majority classes and over sampling of the minority classes have been done.

III. ADVANTAGES

Intrusion detection systems (IDS) are gaining a lot of importance in the present generation. The effort of getting the world closer to each other has been successful with the help of networks but has come along with a lot of disadvantages which mainly includes security of the system through which we transfer our sensitive information across people who are far off in a single click. To fight this problem caused due to lack of enough security intrusion detection systems have been developed which continuously monitors the activity of the user and is good enough to distinguish it from the intruder's activity. It not only detects the intruding activity but is being developed to even prevent such attacks to the system. They are developed to be added on to the software as well as the hardware part of the system and can even be attached separately and can monitor the

activity of the system. They can even be used as Host based intrusion detection (HIDS) where it is installed into individual systems. These systems are even capable of detecting changes like file rewriting or multiple attempts to open a secured file and even accessing the rarely used file frequently, they monitor files, working of routers, key servers, firewall, patterns of malicious content, analyse the types of attacks, shut down access, generate alerts, block IP address and restrict resources. They are trained to use signature-based detection methods or anomaly to detect the threats. They offer a great range of security and visibility to the whole network and organise the process along with the capability the generate a smaller number of false alarms. They are even capable of detecting the specified branch that has to be notified for the particular attack. They have come in great use for domains of military, IT sector and many more.

The fields of Data mining, Machine learning, Artificial intelligence have gained great success in the development of such systems and are continued to put multiple efforts in making them better for further use and generate a smaller number of false alarms. Different types of detection systems are made for different kinds of requirements and the level of security required.Machine Learning and Data Mining techniques have made extensive advancements in their techniques in learning, automation and prediction in the fields of intrusion by brute force and infiltration from within the network. AI for cyber threat detection have already been employed into government agencies, business corporations and private sectors. It saves a lot of human effort and makes things much faster and easier to handle. AI has also extended to Google AI, which is called as the Google Research one of the open-sourced SM3 and an optimizer for large-scale language understanding models like Google's BERT, OpenAI's GPT2. These procedures even improve the cyber security of the systems and prevent the system from future attacks. Artificial Intelligence is very resource intensive and a large community of people have invested in the cyber security community as they work every second for the system which is not possible for any human to do and have any track of any file being written in the system until he opens it for the next set of work he has to carry out. Almost 65% of the organisations think AI is the best mechanism to have for security purposes as it has made way through all the human based work as the growing complexity of the systems have made it impossible to have a continuous watch of multiple things at the same time. AI has made good progress in the parts of accurate and biometric based login techniques, using predictive analysis for intrusion detection, using natural analysis for learning and analysis, using secure access and authentication.

IV. CONCLUSION

In this paper we have discussed about the different techniques which show great potential in detecting intrusion giving an insight of it and the future improvements made to them and that could be made for the better processing and benefits. This paper includes the algorithms of Data Mining, Machine Learning and Artificial Intelligence to meet the requirements. The DM and ML are mostly good for learning and detecting only the learnt threat but AI is proven to be more useful in these requirements as they do apply the same procedure of learning but process and output to a larger extent compared to the other two techniques. It involves the scanning of multiple systems and detailed look into the detection system than the other two. The AI has many open source tools that could help in the detection of such practices and even provides security to an extent. It has many notifying features that are not present otherwise. A large reduce in false positive rates have been noticed when using this and detecting new threats which have not been learnt earlier, in an improved form as compared to the other techniques discussed belonging to other domains. Hence, we can say that this should be encouraged and used in a better manner for future benefits of systems that can detect threats and intrusions. Further improvements like decreasing the false positive rates and further analysis to make them more accurate and be completely independent of human effort should be achieved.

ACKNOWLEDGEMENT

We would like to express our deepest gratitude and respect to our guide Rajeshwari K, Assistant professor Department of ISE, for her guidance, support and encouragement throughout the completion of this paper. Our sincere thanks to The Head of the Department, Dr.M Dakshayini who gave us this opportunity.A special thanks to the Department of ISE at B.M.S College of Engineering for providing all the resources required to facilitate the development of this paper.

REFERENCES

[1] OMC-IDS: At the Cross-Roads of OLAP Miningand Intrusion DetectionHanen Brahmi, Imen Brahmi, and Sadok Ben Yahia.

[2]Network Intrusion Detection using Clustering: A DataMining ApproachS. Sathya Bama, M.S. Irfan Ahmed, A. Saravanan.

[3] Improved Signature Based Intrusion Detection Using Clustering Rule for Decision Tree: Phuong Do, Ho-Seok Kang, Sung-Ryul Kim.

[4] A SURVEY OF DATA MINING TECHNIQUES FOR CYBER SECURITY: Prof. K.P. Barabde, Prof. V. Y. Gaud.

[5]Neural Networks for Intrusion Detection and Its Applications: E. Kesavulu Reddy.

[6] F. Jemili, M. Zaghdoud, and A. Ben, "A framework for an adaptive intrusion detection system using Bayesian network," Intelligence and Security Informatics, IEEE, 2007.

[7]A. Abraham, C. Grosan, and C. Martin-Vide, "Evolutionary design of intrusion detection programs," International Journal of Networks Security, 4 (3), 2007, pp. 328–339.

[8]S. S. Joshi and V. V. Phoha, "Investigating hidden Markov models capabilities in anomaly detection," Proceedings of the 43rd Annual

Southeast Regional Conference, Vol. 1, ACM, 2005, pp. 98–103.

[8] W. Fan, M. Miller, S. Stolfo, W. Lee, and P. Chan, "Using artificial anomalies to detect unknown and known network intrusions," Knowledge and Information Systems, 6 (5)2004, pp. 507–527

[9]Adversarial Reinforcement Learning in a Cyber Security SimulationElderman, Richard; Pater, Leon; Thie, Albert; Drugan, Madalina; Wiering, Marco.

[10]Bolton, R. and D. Hand, Statistical fraud detection: A review. Statistical Science.

[11]Artificial Intelligence in Cyber Security: Amit Rajbanshi, Shuvam Bhimrajka, C. K. Raina.

[12]Artificial Intelligence in CyberDefense: Enn Tyugu.

[13]Intelligent Agents for Intrusion Detection:Guy G. Helmer, Johnny S. K. Wong, Vasant Honavar, and Les Miller.

[14]Anomaly-based intrusion detection using fuzzy rough clustering: W. Chimphlee, A. H. Abdullah, M. N. M. Sap, S. Srinoy, and S. Chimphlee.
[15]Fuzzy network profiling for intrusion detection: J. E. Dickerson, and J. A. Dickerson.
[16] Building lightweight intrusion detection system based on random forest:] D. S. Kim, S. M. Lee, and J. S.Park.
[17]Network Intrusion Detection using Random Forests: J. Zhang, and M. Zulkernine.