INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY (IJCET)

ISSN 0976 – 6367(Print)
ISSN 0976 – 6375(Online)
Volume 4, Issue 3, May-June (2013), pp. 113-122
© IAEME: www.iaeme.com/ijcet.asp
Journal Impact Factor (2013): 6.1302 (Calculated by GISI)
www.jifactor.com

IJCET
©IAEME

# MACHINE LEARNING APPROACH TO ANOMALY DETECTION IN CYBER SECURITY WITH A CASE STUDY OF SPAMMING ATTACK

**Goverdhan Reddy Jidiga[1], Dr. P Sammulal[2]**

[1]Research Scholar, JNTU &
Lecturer in CSE, Dept. of Technical Education, Govt. of Andhra Pradesh, India
[2]Senior Assistant Professor, JNTUCEJ, JNT University Hyderabad, Andhra Pradesh, India

## ABSTRACT

Now the standalone computer and information flow in the internet are sources continues to expose an increasing number of security threats and causes to create a non-recoverable victims with new types of attacks continuously injecting into the network applications. For this to develop a robust, flexible and adaptive security solution is a severe challenge. In this context, anomaly detection technique is an advanced adornment technique to protect data stored in the systems and while flow in the networks against malicious actions. Anomaly detection is an area of information security that has received much attention in recent years applying to most emerging applications. So in this paper we are going to elaborate a latest technique available in machine learning applied to anomaly detection which is used to thwarts the latest attacks created by attackers and here the spam is also a type of anomaly and it is classified as legitimate (ham) or spam. Finally a case study is discussed on latest spamming attacks infected on top web domains and countries in the world motivated by a traditional security ethic are awareness.

**Keywords:** Anomaly, Machine learning, Malware, Phishing, Spam

## I. INTRODUCTION

The Internet is a source of getting information used in our daily life. It is very important tool in the world now used by people in many areas and applications such as personal growth, business, education and many more in the world. The use of internet application such as web and e-mail to communicate with their friends, customers in the

business, but people depending on third parties due to lack of security awareness and unable to know the concepts of information security. This dependence and use of the internet create new a dangerous risk due to increasing attempts from unauthorized third parties to compromise private information for their own benefit. These kinds of situations are part of the cyber crime, therefore essential that all users understand the risks of using internet, the importance of securing their personal information and the consequences if this is not done properly. Today the cyber criminals have such art in their target extremely vulnerable to environment of information technology which includes stealing data and obstructing the operations of the business. Many of the most commonly used systems today are based on signatures based, which are benefited to increasing the false alarms count or conditions that may used indicate may not attack and not caught by existing intrusion detection systems.

Anomaly detection is type of intrusion detection is defined as an intrusion will deviates from normal patterns and the Intrusion detection [27] defined as the process of monitoring the anomaly based events occurring in a computer system or network and analyzing them for signs of intrusions, defined as attempts to bypass the standard security mechanisms of a clean computer or network that are compromise the confidentiality of the valuable data, data integrity, availability and access control of information sources as well as resources. Intrusion detection system (IDS) [13] is a combination of software and hardware that attempts to perform protection to normal users and system resources from threats. There were numerous attacks on software systems result in a process execution or human coding mistakes deviating from its normal behavior [2], all these prominent examples include a malware related code injection attacks on internet server's processes. Up to now we have seen significant amount of research to detect such attacks through monitoring the behavior of the suspected process and comparing that behavior to a model [3] consist of normal behavior collected from past experiences. These are also called anomaly detection techniques because in compare to signature based detection which deviates from the normal behavior are taken as indications of anomalies. Anomaly detection may be divided into static and dynamic anomaly detection or divided based on nature of behavior those are statistical and knowledge based and generic machine learning based techniques.

## II. ANOMALY DETECTION

### 2.1 Introduction

In this the basic unit of finding abnormal behavior is identified as an anomaly. Anomaly [7, 27] is a pattern in the data that does not conform to the expected behavior. Anomaly detection is monitors program executions and detects anomalous program behaviors through reverse analysis of executable program including a critical behavior monitoring points can be extracted from binary code sequences [2] and memory space. Most of the available IDSs are signature based, so such systems are not used to find the frequent rule based attacks, unknown attacks and updates. The existing systems even if it is designed by traditional and advanced anomaly detection techniques are not observing real world anomalies like emerging cyber threats, cyber intrusions, credit card frauds...etc. The anomalies occur relatively infrequently and their consequences can be quite dramatic, negative sense in the running of applications.

### 2.2 Why Anomaly Detection

The main advantage of anomaly detection technique is their probable to detect previously unseen anomalies events. The security breaches are very common now in the society and organizations fail to take effective measures. When it comes to new technologies organizations have needed to move quickly, but they are not responding fast especially in critical infrastructure are worst [19]. The business and personal use they have forced organizations to urgently implement policies that address the risks associated with an evolving array of emerging technologies.

Also the organizations are protecting how they guard to their data of the employees and their customers like service based organizations in the world with new cyber attacks. Today cyber attacks are common in the public banking sector, health organizations, defense, and service sector , so organizations are need to give training and guidelines, policy adjustments, stepping up awareness programs. So our aim is to prepare for an effective solutions working online and on the fly counter action is required to avoid the cyber criminals, viruses, malware and botnets shown. Now experts need to not only consider how they can occur and use powerful analytics to detect security events but also realize to aware of dynamic threats caused by malicious events.

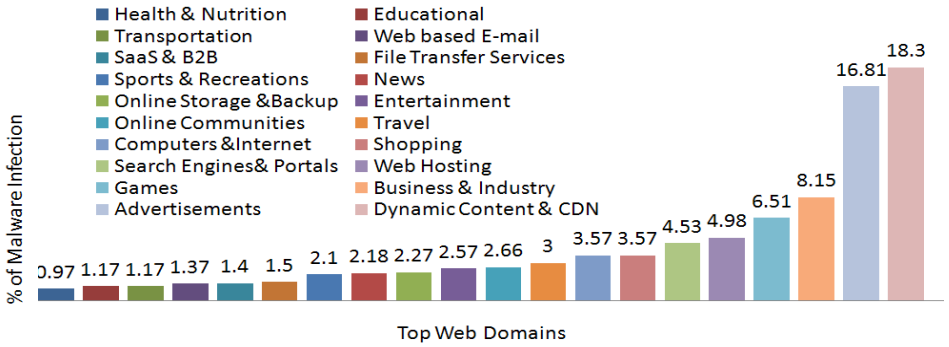


**Fig.1.** Top web domains infection over worldwide up to 2012. (Source: Cisco security survey Report)

According to information security survey [19, 20, 21] up to 2012 the threats and malware infections on the top domains in the world shown in Fig.1. To overcome these threats we use the anomaly detection with usage of machine learning approaches [27]. The main uses for Anomaly Detection are detect precedent attack behavior, zero day attack detection, insider threat detection, and validate or assist with signature data. The major advantages of anomaly detection in this paper are mail fraud and credit card fraud detection. Anomaly detection (AD) systems have some advantages [25, 27]. First AD have the capability to detect insider attacks like someone using a stolen account starts performing actions that are outside the normal user profile, an anomaly detection system generates an alarm. Second, AD is based on customized profiles, it is very difficult for an attacker to know with certainty what activity it can carry out without setting off an alarm. Third, an anomaly detection system has the ability to detect previously unknown attacks.

## III. RELATED WORK

Most of the anomaly intrusion detection systems are signature based and fundamental statistics or knowledge based, but these are all suitable in some applications and not suitable today in advanced technical concepts. Now we discussed some related work on old and new one is based on machine learning discussed in next paragraph. Anderson is the first person elaborated the intrusion concept in security and he developed model [1] threats are classified as masqueraders, misfeasors and clandestine users. The Anderson model is good initially, but now it is not suitable. Denning proposed several models for Intrusion Detection System (IDS) development based on statistics, Markov chains, time-series, etc [4]. In Denning model, user's behavior that deviates sufficiently from the normal behavior is considered anomalous. The system they developed was only used offline using previously collected data and is not suitable to detect the cyber attacks.

Machine learning based work: In 2000, Valdes [9] developed an anomaly based intrusion detection system that employed naive Bayesian network to perform intrusion detecting on traffic bursts. In 2003, Kruegel [24] proposed a multisensory fusion approach using Bayesian classifier for classification and suppression of false alarms that the outputs of different IDS sensors were aggregated to produce single alarm. In the same year, Shyu [10] proposed an anomaly based intrusion detection scheme using principal components analysis (PCA), where PCA was applied to reduce the dimensionality of the audit data and arrive at a classifier that is a function of the principal components. In [17,18] proposed an anomaly based intrusion detection using hidden markov models(HMM) that computes the sample likelihood of an observed sequence using the forward or backward algorithm for identifying anomalous behavior from normal behaviors. Lee [6, 11] proposed classification based anomaly detection using inductive rules to characterize sequences occurring in normal data. In 2000, Dickerson [12] developed the Fuzzy Intrusion Recognition Engine using fuzzy logic that process the network input data and generate fuzzy sets. The other techniques such as Naive Bayes theorem, SVM, ANN, Regression technique, Artificial Immune system, Lazy learning, Rough set theory, k-NN, Genetic algorithm...Etc. So therefore, the primary and most important challenge is we needs to be develop the on the fly countermeasures and effective strategies to reduce the high rate of false alarms by using of machine learning rules.

## IV. MACHINE LEARNING

In this paper we concentrated on machine learning [15,25] techniques , because it use strong statistical foundations to enhancing the dynamic and accurate learning that gives better accuracy, small false alarm rates, learned detectors use a more compact representation, possible performance improvements, ability to detect novelty, protection against zero-day exploits, faster incident response times.etc. In the machine learning different novel contributions of techniques include taxonomy of different types of attacks on systems, a variety of defenses against those attacks.

In the fig.2 the anomaly detection taxonomy[7,13] is given, it is based on classification of anomaly detection which is purely based foundational work done past authors of intrusion detection systems models and today performance based machine learning approaches. In this taxonomy the anomaly detection is based on machine learning and data mining approaches. The machine learning use strong statistical foundations to enhancing the dynamic and accurate learning that gives better accuracy, small false alarm rates.

### 4.1 Proposed Model

In fig.3 the machine learning based AD (Anomaly Detection) is used and prototype is given with preprocessing data. In AD (Anomaly Detection) prototype model the audit data collection module is used in the data collection phase. The data collected in this phase is analyzed by the anomaly detection algorithm to find traces of suspicious activity. The source of the data can be host/network activity logs, command-based logs, application-based logs, etc. audit data in intrusion detection systems store the audit data either indefinitely or for a sufficiently long time for later reference. The volume of data is often exceedingly large, so persistent database is maintained.

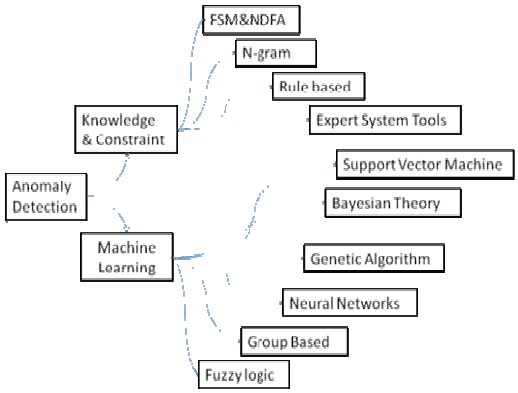


**Fig.2.** Taxonomy of AD based on Machine learning (Right Branch).

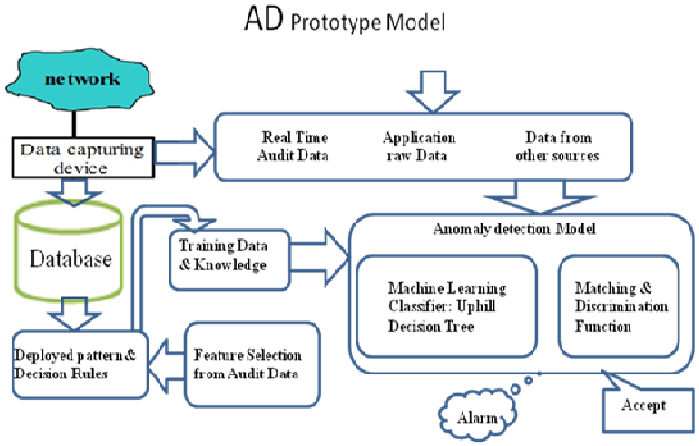


**Fig.3.** Anomaly Detection (AD) Prototype Model with Machine learning

In Anomaly detection due to the complex and dynamic properties of anomaly behaviors, machine learning and data mining techniques generally mixed to optimize the performance of anomaly detection systems to finding specific point anomalies or range anomalies at moment of time. We give an efficient algorithm for provably learning uphill decision tree with extension adornments of existing multi-way decision tree algorithm. In fig.3 the machine learning is decision tree algorithm is considered first and later it is extended

to uphill decision tree also called regression tree. The processing element must frequently store intermediate results such as information about partially fulfilled anomalies. The model contains a logic taken from uphill decision tree detect the anomalies by raising notified alarms.

### 4.2 Proposed Machine Learning Algorithm : A Uphill Decision Tree (UDT)

The Decision tree (DT) learning [16, 27] is a type of machine learning algorithm used in many application of information security in previous research. The decision tree (DT) is very powerful and popular data mining algorithm for decision-making and classification problems. It has been using in many real life applications and can be constructed from large size of data collected from attributes defined in the problem. A DT has three components: nodes, leaves, and edges. Each node is labeled with an attribute by which the data is to be partitioned. Each node has a number of edges, which are labeled according to possible values of the attribute. An edge connects either two nodes or a node and a leaf. Today it is olden and not effecting in current cyber attacks. So the extension of this is an uphill decision tree.

A decision tree with real values at the leaves is called an uphill decision tree if the values on the leaves are non-decreasing in order from left to right. An Uphill decision tree is similarly a tree structured solution in which a constant or a relatively simple regression model is fitted to the data in each partition. In this algorithm we considered the data collected for e-mail to filtering whether it is phishing or spam.

Phishing is a type of internet fraud deployed to steal confidential financial information that includes theft of net banking passwords, corporate secrets, credit card numbers, financial status, bank account details and other valuable information and spam [19] is anonymous, unsolicited bulk email diverting the cybercitizen's minds to use their services and products etc. phishing is also type of spam. The total number of spam emails are increasing day to day, up to 2012 the top domains and areas in the world suffering with cyber threats like DOS, DDOS,SQL Injection, spamming attacks , phishing attacks and others. In that most of them are spam , phishing attacks affected on online banking, Online purchasing (PayPal, Amazon, eBay, etc.), Social media (Face book, Twitter, blogs, etc.) in all corners.

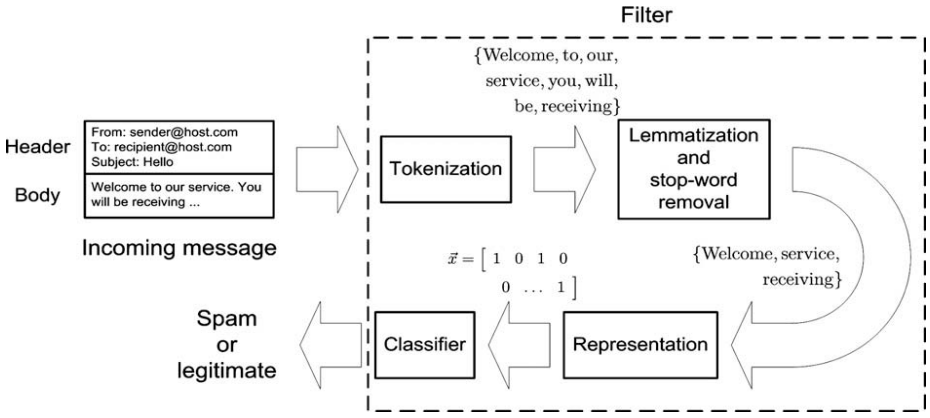


**Fig.4.** Spam Filter for e-mail verification to estimate the Spam attack.
(Source: www.elsevier.com/locate/eswa)

The semantics or components of e-mail are like domain, class, frequency, link, URL, IP address, script, validation, port address, dot, images, no. of ports valid or not, link valid or not, mismatching .etc available. Based on that we can estimate the mail or websites are legitimate or spam by using a pre-determined set of rules designed in the construction of uphill decision tree.

In fig.4 we have applied a concept of semantics for e-mail verification whether the e-mail is come from legitimate organizations or not. The mail semantics are compiled by filter one by one and find the some unknown semantics are encountered that we compare with original semantics of e-mail. In this we take an example-1 of e-mails is registrar@jntuh.ac.in. Here the semantics are shown in order in fig.5.

If(domain=TRUE&e-mail-has=HTML&script=Java_script&Validation=TRUE&server=authenticate)
then P (e-mail Semantics) = "90%". (This value depending on other semantics also)

In this e-mail, if all semantics are correct and verified by decision tree including URL, no. of dots in URL, IP address and port of application then we can probably identify that mail come from authorized party.

Example-2: VISA card related mail from bank contains unknown hits like "dear valued customer" shown in fig.5, But in original mail from bank is not contains semantics like "dear valued customer"

Here if (domain=TRUE &e-mail-has="unknown hits") then P (e-mail Semantics) = "20%". (This value depends on other semantics also). In this e-mail, first semantic is correct and verified by decision tree and semantic at node-3 has unknown hits may not included in semantics of VISA mail including URL, no. of dots in URL, IP address and port of application then we can probably identify that mail come from unauthorized party.

Here if (domain=TRUE &e-mail-has="unknown hits") then P (e-mail Semantics) = "20%". (This value depends on other semantics also). In this e-mail, first semantic is correct and verified by decision tree and semantic at node-3 has unknown hits may not included in semantics of VISA mail including URL, no. of dots in URL, IP address and port of application then we can probably identify that mail come from unauthorized party consider as spam.

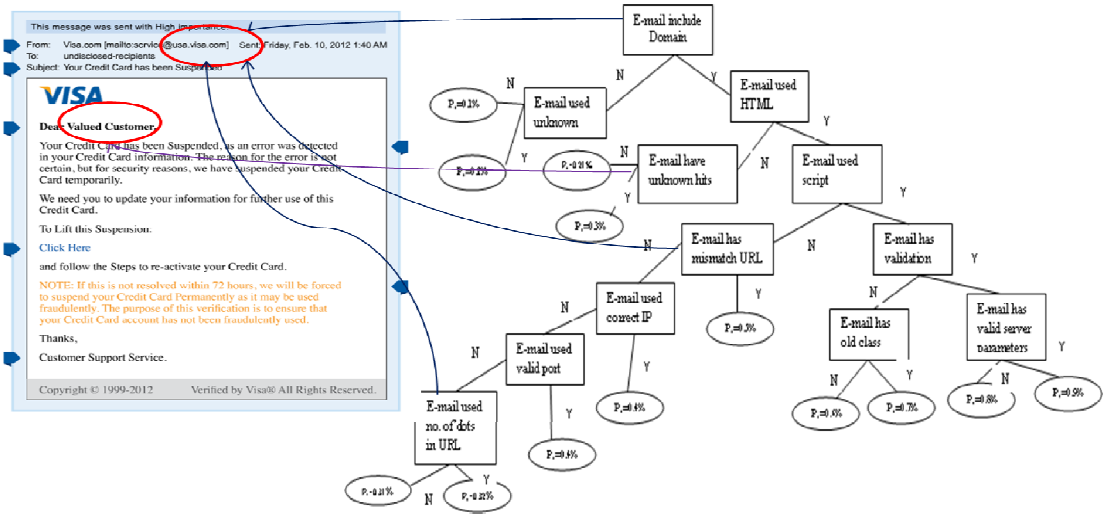


**Fig.5.** Spam attack e-mail verification by semantics of VISA (Source: VISA Card Security)

### 4.3 Ethical Solution to Spam : E-Awareness with Case Study

The basic E-awareness [21] is a process of keeping people in continuous attention of security to save information. We know that e-awareness also a part information security ethics to thwart a most vulnerabilities as "security awareness is better than prevention and prevention is better than detection". This was an ethic concept applied in all kinds of human life applications to survive in the nature. The people who are expert in security aspects to thwart the security deficiencies, eligible to train all users of information technology to identify and report the all kind of suspicious activities in their electronic environment. Now it is essential that each of us take responsibility and understand our role in securing cyberspace. ACM Report [26] given the countermeasures on spam and phishing is creating awareness and train end users to proactively recognize and avoid spam attacks (ethical and very popular approach). The solutions are motivating people to be secure, micro games designed to teach people about phishing and embedded training.

The Spam attacks [19-22] are very dangerous position in India shown in fig.6. From 2009 onwards it is occupying 1st rank in the world and maximum spam shared by BRIC (Brazil, Russia, India, China) countries only and USA itself controlling the spam by taking necessary actions like conducting e-awareness programs frequently. The Vietnam occupied 3rd rank in 2011 and 2013 respectively, and Russia also a victim by 2nd rank in 2011.

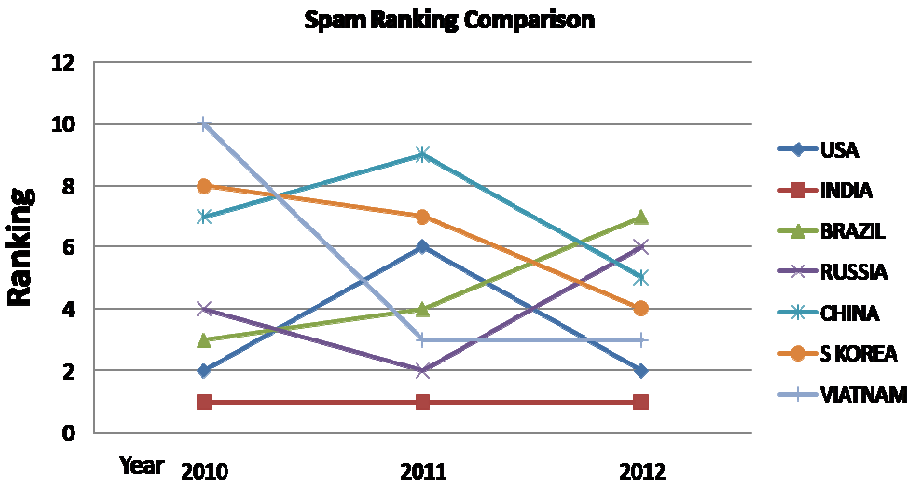


**Fig.6.** Spam attack Ranking

What are the reasons for increasing phishing and spam [20] in India are:

1) Lack of awareness, education and responsibility, Greediness in earning of easy money.
2) Lack of use of new technology, using of poor technology and use of pirated software.
3) High unemployment, illiterate, ignorance, population and competition in the market.
4) Lack of government support, policy constraints, coordination, law.
5) May fast economy development and technology using.
6) Huge growth in the usage of Mobiles, System sharing, Rate of using internet.

## V. CONCLUSION

In this paper we used an uphill decision tree machine learning approach to anomaly detection is temporarily a solution and applying to spam attacks. Now spam and phishing will persist in any electronic medium pursue a problem that can never truly be solved. In this nest better we can work on always to preventing, detecting the spam, and responding to this e-awareness. Finally in this paper we present a case study on spam attack based on the awareness model and today the machine learning is only approach encouraged by well known scientists in the field of security. So this will give concepts and motivates to you a do further research and also hope that this work to be true at our knowledge.

## REFERENCES

[1] J.P.Anderson,"Computer security threat monitoring and surveillance," James P Anderson Co.,Fort Washington,Pennsylvania, USA, Technical Report 98–17, April 1980.

[2] H. H. Feng, Oleg M. Kolesnikov, P. Fogla, Wenke Lee, and Weibo Gong, "Anomaly Detection Using Call Stack Information"IEEE Symposium on Security and Privacy'2003, CA, Issue Date: 11-14 May2003 pp: 62-75 ISSN: 1081-6011 Print ISBN: 0-7695-1940-7.

[3] D. Wagner,D.Dean,"Intrusion Detection via Static Analysis",IEEE Symposium on Security and Privacy, Oakland, CA, 2001.

[4] D.E.Denning "An intrusion detection model" In IEEE Transactions on Software Engineering, CA,1987. IEEE Computer Society Press.

[5] Mukkamala,J.Gagnon,andS. Jajodia."Integrating data mining techniques with intrusion detection methods" Research Advances in Information Systems Security, Kluwer Publishers, Boston, MA. 33-46,2000.

[6] W.Lee, ChanP.K, Eskin, E WeiFan, Miller M. S.Zhang "Realtime datamining based intrusion detection" IEEE DARPA information Conference 2001,DISCEX'01,Proceedings IssueDate: 2001 page(s):89-100vol.1 12 Jun2001-14 Jun 2001 Print ISBN:0-7695-1212-7.

[7] S.Axelsson,"IDS: A Survey and Taxonomy," Chalmers University, Technical Report 99-15,March 2000.

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

[9] A. Valdes and K. Skinner, "Adaptive Model-based Monitoring for Cyber Attack Detection," in Recent Advances in Intrusion Detection Toulouse, France, 2000, pp. 80-92.

[10] M.L.Shyu,S.C.Chen,K.Sarinnapakorn,and L.Chang,"A Novel Anomaly Detection Scheme Based on PCA Classifier," in IEEE Foundations and New Directions of DataMining Workshop, Florida,USA,2003,pp.172-179.

[11] W. Lee and S. J. Stolfo, "Data mining approaches for intrusion detection," in 7th USENIX Security Symposium (SECURITY-98), Berkeley, CA, USA, 1998, pp. 79--94.

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

[13] L.Ertoz, E.Eilertson, A.Lazarevic, P.N. Tan, V. Kumar, J. Srivastava, and P. Dokas, "The MINDS - Minnesota Intrusion Detection System", in Next Generation Data Mining Boston: MIT Press, 2004.
[14] S.Mukkamala, G.I.Janoski, and A.H.Sung."Intrusion Detection Using Support Vector Machines",Proceedings of the High Performance Computing Symposium- HPC 2002, pp 178-183, San Diego, April 2002.
[15] T. Lane and C. E. Brodley. "An Application of Machine Learning to Anomaly Detection", Proceedings of the 20th National Information Systems Security Conference, pp 366-377, Baltimore, MD. Oct. 1997.
[16] Quinlan, J. Ross, "Induction of Decision Trees," Machine Learning, 1:81{106, 1986. Reprinted in Shavlik, J. and Dietterich, T.,Readings in Machine Learning, San Francisco: Morgan Kaufmann, 1990, pp. 57-69.
[17] Ghahramani Z,"An introduction to hidden markov models and bayesian networks". HMM: applications in computer vision, pages 9–42, 2002.
[18] HaiTao H,XiaoNan L,"A novel HMM-based approach to anomaly detection", Journal of Information and Computational Science 1 (3) (2004) 91–94.
[19] An Article and report on spam http://www.antiphishing.org/
[20] An Article and report on spam http://india.emc.com
[21] An Article and report on spam http://www.rsa.com
[22] An Article and report on spam http://www.cisco.com
[23] A. Ghosh and A. Schwartzbard, " A study in using neural networks for anomaly and misuse detection", 8th USENIX Security Symposium, pp. 141-151, 1999.
[24] C. Kruegel, D. Mutz, W. Robertson and F. Valeur, "Bayesian Event Classification for Intrusion Detection," in 19th Annual Computer Security Applications Conference, Las Vegas, NV, 2003.
[25] L.Breiman,"Random Forests,"Machine Learning,vol. 45, pp. 5-32, 2001.
[26] By Jason Hong, "article on The State of Phishing Attacks" Communications of the ACM, Vol. 55 No. 1, Pages 74-81.
[27] A. Patcha, J-M. Park, "An Overview of Anomaly Detection Techniques: Existing Solutions and Latest Technological Trends", Computer Networks(2007).
[28] Mr. Sachin J.Pukale Mr. M. K.Chavan, "A Review of Anomaly Based Intrusions Detection in Multi-Tier Web Applications", International Journal of Computer Engineering & Technology (IJCET), Volume 3, Issue 3, 2012, pp. 233 - 244, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.
[29] C.R. Cyril Anthoni and Dr. A. Christy, "Integration Of Feature Sets With Machine Learning Techniques For Spam Filtering", International Journal of Computer Engineering & Technology (IJCET), Volume 2, Issue 1, 2011, pp. 47 - 52, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.
[30] A.Edwinrobert and Dr.M.Hemalatha, "Behavioral and Performance Analysis Model for Malware Detection Techniques", International Journal of Computer Engineering & Technology (IJCET), Volume 4, Issue 1, 2013, pp. 141 - 151, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.