COMPARATIVE ANALYSIS OF ANOMALY BASED WEB ATTACK DETECTION METHODS

Achin Jain and Vanita Jain

BharatiVidyapeeth College of Engineering, New Delhi, India

ABSTRACT

In the present scenario, protection of websites from web-based attacks is a great challenge due to the bad intention of the malicious user over the Internet. Researchers are trying to find the optimum solution to prevent these web attack activities. There are several techniques available to prevent the web attacks from happening like firewalls, but most of the firewall is not designed to prevent the attack against the websites. Moreover, firewalls mostly work on signature-based detection method. In this paper, we have analyzed different anomaly-based detection methods for the detection of web-based attacks initiated by malicious users. Working of these methods is in a different direction to the signature-based detection method which only detects the web-based attacks for which a signature has been previously created. In this paper, we have introduced two methods: Attribute Length Method (ALM) and Attribute Character Distribution Method (ACDM) which is based on attribute values. Further, we have done the mathematical analysis of three different web attacks and compare their False Accept Rate (FAR) results for both the methods. Results analysis reveals that ALM is more efficient method than ACDM in the detection of web-based attacks.

KEYWORDS

Web-Based Attacks, Anomaly Based Detection, Attribute Length Method, Attribute Character Distribution Method, False Accept Rate.

1. INTRODUCTION

Web traffic is increasing day by day and, therefore, there are large numbers of attacks occurring on the websites. In the present World Wide Web traffic scenarios, where everything is connected to the Internet and everyone in the industry succeeds for their success on the strong presence in the online market. The attack on websites is on the increase whether it has defaced the image of the competitor or to hack website just for fun. But ultimately it is the loss of the owner of the website. Attacks on business related websites are more threatening than on personal websites because business websites contain all the financial information about the industry, employee details and lot more things. Web applications are becoming more and more professional and business oriented in all sorts of areas ranging from social media to the online shopping world. The result of such rapid increase in the popularity of the website brings both positive and negative sides to the picture. With the increase in traffic revenue of the business attackers also comes to know about the site and they like to attack sites with heavy traffic, therefore, the majority of the visitors are affected. The effects of the attacks are not only felt during the attack but also they leave after attack effects also like identity disclosure, hijacking of sensitive data information and unauthorized access to the information. Therefore, apart from running the website it is necessary and important task to protect web applications and adopt suitable security methods.

DOI: 10.5121/ijci.2015.4602

One of the most promising solutions proposed for detecting web-based attacks is the use of an anomaly-based intrusion detection system (IDS). In this paper, we have considered various anomaly detection methods like "Attribute Length" and "Attribute Character Distribution" proposed in [1]. The anomaly-based detection method is preferable over the signature detection method because in the later detection of attacks depends only on the database of earlier occurred attacks. If there is a new attack with new payload and functionality then the signature method will not be able to detect the attack. Anomaly-based web attack detection method is a method for detecting web attacks and malicious activities by monitoring system activities and classifying the activities as either normal or malicious. The classification is based on rules that are generated by working in the normal training dataset and the setting of the threshold value which is used in the comparison process. This method works in the different direction to the signature-based detection method which only detect the web-based attacks for which a signature has been previously created. However, the major problem in anomaly-based approaches is that they tend to suffer from a high rate of false alerts [2, 3].

The rest of the paper is organized as follows: Section 2 Related Works carried out by researchers in web attack detection. Section 3 gives a brief overview of the different types of web attack on web applications. In the section, 4 descriptions of anomaly-based web attack detection methods are given in detail. Section 5 explains the proposed work. Experimental work and result analysis of the proposed work is carried out in section 6. Finally, section 6 concludes this paper.

2. RELATED WORK

Bolzoni, Etalle and Hartel in [5] proposed a novel Anomaly-based NIDS system known as POSEIDON. In this system the researchers modified original PAYL method and used unsupervised classification technique for preprocessing task. In article [6] Shyu, Chen et.al used robust principal component to classify Intrusion Detection problems. This modified approach is better from its original method in its ability to differentiate the anomalies from its normal instances in tens of tremendous values of various correlation structures.

In the past couple of years researchers are working very hard to find an optimum solution for the detection of web based attacks. In the effort Kruegel et.al in [7] proposed an Intrusion Detection System (IDS) based on Anomaly Detection techniques for the detection and prevention of web attacks. In [8] Noble and Cook used graph-based statistics techniques for the detection of unusual patterns. The authors used the concept of conditional entropy to calculate the regularity of the graph leading to the detection of successful anomaly detection. The method is useful for ruling out the anomalies in the set of web data.

Maxion and Tan in [9] have proposed a metric for the characterization formation in data environments and found that fundamental structure greatly affect the probabilistic discovery. In [10] Tapiador, Teodoro, and Verdejo used HTTP request monitoring to propose new approach based on Markovian model for the detection of web attacks. Gomez and Dasgupta in [11] created a fuzzy system using Genetic Algorithm that is able to detect anomalies and intrusions. The authors in this approach design a classification system based on fuzzy logic to divide normal request from malicious.

Tapiadoret. al in [12] used Markov chains model to propose a new solution for the identification of attacks carried out over HTTP traffic. Mehta and Jamwal in [13] used QualysGuard WAS tool to minimize the web vulnerabilities. In the paper authors focused on Cross Site Scripting (XSS) code which is proven to be the main source of web based attacks. In the article they have shown

that QualysGuard WAS tool is the best solution to satisfy client and server side security requirements.
In [20], Patil et.al have proposed an effective concept using ID3 algorithm for detection of web based attacks, using data set from 'SmarSniff' tool. In the experimental results shown by the author shows that improved algorithm is effective in decrease the data amount and reduce the impact of data with poor quality.Snigdha et.al in [21] have implemented the Attribute Length of Web Attack Detection method using Java Programming language. After training the model they have tested the solution on real time data and from the results shown that attribute length method is most suited in detecting web based attacks where parameters are fixed size tokens.

In [22] the authors have presented a framework for generating synthetic datasets with normal and malicious data for web applications across multiple layers simultaneously. For the experimental work they build a prototype data generator using to generate nine datasets with data logged on four layers: network, file accesses, system calls, and database simultaneously and tested 19 security controls.In [23] the authors have proposed a fraud detection system that uses different anomaly detection techniques to predict computer intrusion attacks in e-commerce web applications.Han et. al in article [24] used Request Length Module using Regex Pattern Analysis for the detection of three types of attacks namely SQl injection, XSS and Directory Traversal Attack.

3. WEB BASED ATTACKS

An attack is defined as the unwanted intrusion to website resources. Attackers can attack the website for different reasons and purposes. If the intention of the attacker is just to monitor the traffic and information flowing in and out the web server then it is termed as "Passive Attack". When the attacker tries to modify or destroy the web resources then it is termed as "Active Attack". There are many solutions available to counter these attacks such as firewall etc. But these solutions are not always enough to protect the users from being attacked. As a result users are vulnerable to exploitations while performing basic functionalities (e.g., login) [4, 5]. In this paper, we have restricted our research work to three well-known vulnerabilities explained in the Open Web Application Security Project (OWASP) [14]. For experimental work, we have taken three different types of the active attack i.e. Cross Site Scripting Attack (XSS), Buffer Overflow Attack, and Path Traversal Attack. These are described below in brief.

3.1. Cross Site Scripting Attack (XSS)

Cross Site Scripting (XSS) attack refers to a range of attacks in which the attacker injects malicious code mostly JavaScript into a web application [15, 16]. According to [17] more than 60% of websites are vulnerable to XSS attacks. The XSS attack can take on many forms like for example execution of JavaScript commands on the web server by passing of script in place on the parameter value in the URL. Here, we have used different XSS commands from OWASP for the experimental work on attribute length and attribute character distribution method. A simple example of XSS attack is shown below:

$$<img""><script>alert("XSS")</script>">$$

3.2. Buffer Overflow Attack (BOA)

In this attack, the attacker tries to exploit the very common vulnerability in the web server of not validating the input properly. A buffer overflow occurs during program execution when a fixed-

size buffer has had too many data copied into it. This causes the data to overwrite into adjacent memory locations and depending on what is stored in it. The behaviour of the program itself might be affected [18]. Execution of this attack can have many consequences, but the most one is non-availability of the web application. In this work, we have done the padding of the training data set parameter value to increase the length so that we can test out both anomaly-based detection methods. A simple example of BOA is shown below:

$$http://localhost/Test.php?$$
$$yourname = achinachinachi$$
$$nachaicnainciancainininivr$$
$$ivnrivnrivrnvirnvivnivrnvir$$
$$nvirnvrvnrvirnvrivnrivnriv$$

3.3. Path Traversal Attack (PTA)

PTA exploits the vulnerability which is related to the web server path directory. The intention of using attack is to access the files and directories that are not meant to be accessed by unauthorized persons. This attack works on applications that take user input and use it in a "path" that is used to access a file system [19]. Modus Operandi of this particular attack is to insert special characters in the URL to try to modify the meaning of the path, the result can be misbehaving of the application and may allow accessing the private resource to the attacker. '.. /' Character is the most basic special sequence that is used in PTA to alter the location of the request. A simple example of PTA is shown below:

$$item = ../../../etc/passwd$$

4. TECHNICAL APPROACH OF ANOMALY BASED DETECTION METHODS

Anomaly-based web attack detection method is a method for detection web attacks and malicious activity by monitoring system activity and categorizing the activity as either normal or malicious. The Categorization is based on rules that are generated by working in the normal training dataset and the setting of the threshold value which is used in the comparison process. The anomaly detection process uses a number of different models to identify anomalous entries within a set of input request Ur associated with a program r. All the three attacks discussed above are analyzed with attribute length and attribute character distribution method proposed in [1].

4.1. Attribute Length Method (ALM)

This method is based on the fact that the length of the attribute in the query string can be used to detect malicious request. This method is most effective in cases where the parameter is set as fixed-size tokens. In normal cases when the parameter value is passed for a certain case like username (which we have used here for evaluation) does not deviate much. However, in case of malicious input attackers tends to increase the parameter value. For example in BOA attackers increase the length by padding with additional characters. The other two attacks XSS and PTA which we have evaluated in this work can also be detected by the ALM.

4.2 Attribute Character Distribution Method

In this method frequency distribution of the characters is analyzed to calculate the Ideal Character Distribution (ICD). This approach is based on the observation that attributes have regular structure are mostly human readable but in case of attack that send large data (BOA), different character distribution is obtained due to the padding of the extra characters. This method can also be used to detect the XSS and PTA as we have carried out in the experimental work.

5. PROBLEM DEFINITION

This work mainly focuses on the analytical evaluation of three above discussed web-based attacks, i.e. XSS, BOA and PTA attacks. In this work, we have evaluated the detection of web-based attacks using "Attribute Length" and "Attribute Character Distribution" anomaly detection techniques. We have divided the entire work in two different phases: the training phase and testing phase. In the training phase which helps us in finding out the threshold value which will be used as measurement to detect malicious and normal query. Kruegel and Vigna utilize different models in [1] to detect the web attacks in HTTP request that contain a query section. Also, we have computed the probability value for both the methods in the training phase and then investigate the probability value in case of malicious parameter for the testing phase. By analyzing the deviation, we have carried out the detection process of web-based attacks. We will describe both the phases in details that we have done for our experimental work.

5.1. Attribute Length Method (ALM): Training Phase Analysis

We have captured the length of the parameter (username) for each HTTP request that is passed during the training phase. Compute the mean length by adding all the length and divided by the total number of parameters then we have calculated the threshold probability of using the equation given below:

$$P = \frac{\sigma^2}{(L-\mu)^2}$$

Where L ($L \geq \mu$) is the length of the string parameter and μ is the mean length of all the string parameters during the training phase and $\sigma 2$ is the variance calculated for all the inputs. The experimental values are shown in Table 1.

Table 1. Parameters values

Parameters	**Values**
Variance($\sigma 2$)	11.7313
Mean Length(μ)	8.68966
Length(L) of Parameter 'Username'	Vary in different query input
Threshold Probability	0.79

Threshold probability (Pth) is 0.79 for the rest of the work in Attribute Length Method.

5.2 Attribute Length Method (ALM): Testing Phase

After computing Pth, we have followed the same process for three different web attacks discussed above. Probability values for the above mentioned attacks are shown in Table 2.

Table 2: Probability Value (Malicious Data)

S. No.	Threshold Probability Value	Probability value in BOA	Probability value in XSS	Probability value in PTA
1	0.79	0.631	0.008	0.035
2	0.79	0.416	0.015	0.416
3	0.79	0.295	0.020	0.039
4	0.79	0.219	0.022	0.170
5	0.79	0.170	0.012	0.219
6	0.79	0.135	0.005	0.631
7	0.79	0.110	0.014	0.057
8	0.79	0.092	0.015	0.031
9	0.79	0.077	0.022	0.950
10	0.79	0.066	0.024	0.950
11	0.79	0.057	0.950	0.039
12	0.79	0.050	0.631	0.950
13	0.79	0.044	0.950	0.950
14	0.79	0.039	0.018	0.295
15	0.79	0.035	0.017	0.631
16	0.79	0.031	0.016	0.024
17	0.79	0.028	0.009	0.018
18	0.79	0.026	0.009	0.135
19	0.79	0.024	0.010	0.219
20	0.79	0.022	0.008	0.018
21	0.79	0.020	0.950	0.017
22	0.79	0.018	0.219	0.015
23	0.79	0.017	0.110	0.014
24	0.79	0.016	0.077	0.066
25	0.79	0.015	0.050	0.031

5.3 ALM: Result Analysis

In this section, we have compared the observed probability values with the threshold probability value (Pth) experimentally for testing phase as shown in the fig. 1. The threshold probability value is shown with a blue line in the graph and the value that lies behind this trend line are the values in web-based attacks that are detected and the value above blue line are the attacks that are not detected by the attribute length method.

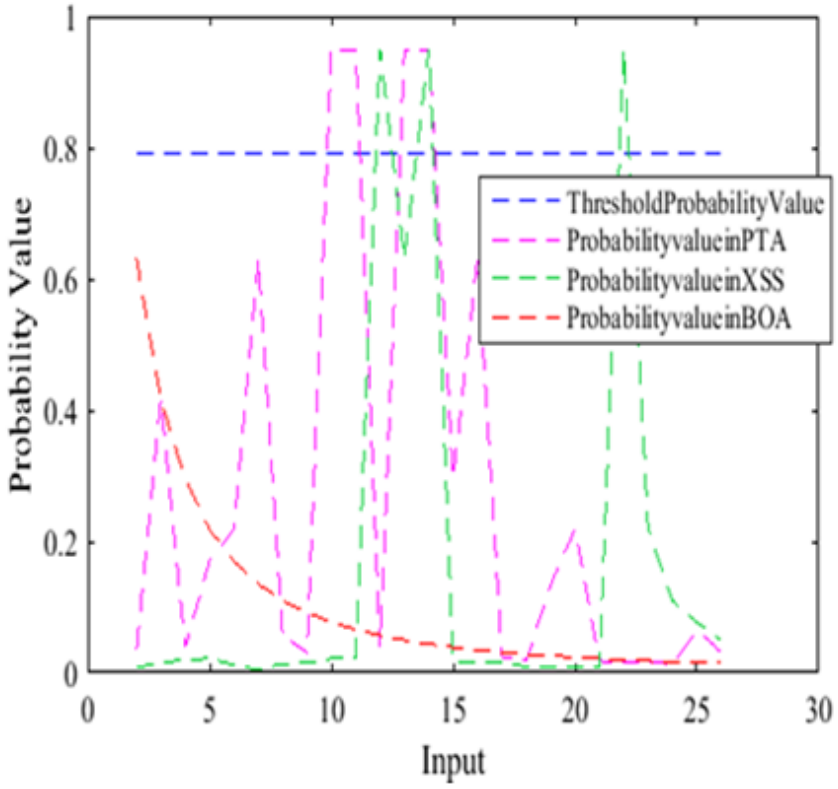

Fig. 1: Threshold probability value and web-based attacks probability value

The False Accept Rate (FAR) for ALM is calculated using the equation given in the box and results are shown in the Table 3.

$$FAR = \frac{Number\ of\ malicious\ query\ accepted}{Total\ number\ of\ queries}$$

Table 3: FAR calculation in ALM

Variables	**Values**
Total Number of Queries	75
No. of Malicious Queries accepted	7
FAR	9.3%

From above table 3, it is clear that only 9.3% web-based attacks are not successfully detected in the area above the red line in the graph. It means that attribute length method successfully detected 90.7% web attack in the area below the red line as shown in the graph.

5.4 Attribute Character Distribution Method (ACDM): Training Phase Analysis

In this work, we have introduced the training model for ACDM. In this model each HTTP request that is passed during the training phase. We have computed the character distribution of each alphabet from A-Z. However, in real situation all the characters including special character may be considered but for this work we have considered only the alphabets for constructing the

training data set. After finding out the distribution of each character we have divided them into six different bins that are shown in the table 4, below.

Table 4: Distribution of characters in six bins

Bin Generated	Characters in the Bin
Bin1	{A, Z, X, W}
Bin2	{I, Q, P, G}
Bin3	{N, B, F, J}
Bin4	{H, M, S, D}
Bin5	{R, T, V, Y, C}
Bin6	{E, O, L, K, U}

After dividing the characters into six bins, we have computed the threshold probability value (Pth) for training data using the equation given below.

$$X^2 = \frac{\sum (O_i - E_i)^2}{E_i}$$

Where, Oi is the observed frequency of characters of same bin and Ei is the expected frequency of the bin which is calculated by the formula given below.

$$E_i = L\,x\,F.D(bin)$$

Where, L is the length of the query parameter and F. D (bin) is the frequency distribution of the bin in which the character lies. The probability is calculated using the chi square table with degree of freedom as 5. Threshold probability Pth computed in the training phase is 0.041.

5.5 ACDM: Testing Phase Analysis

After computing the Pth, we have carried out the same work for three web-based attacks that are discussed above. Probability values for different malicious parameter values in BOA, XSS and PTA are shown in table 5.

Table 5. Threshold probability values against probability in attacks for malicious parameter value

S. No.	Threshold Probability Value	Probability in BOA attack	Probability in XSS attack	Probability in PTA attack
1	0.040	0.052	0.044	0.047
2	0.040	0.016	0.012	0.010
3	0.040	0.020	0.018	0.035
4	0.040	0.040	0.040	0.022
5	0.040	0.019	0.020	0.035
6	0.040	0.050	0.056	0.040
7	0.040	0.004	0.002	0.026
8	0.040	0.016	0.014	0.046
9	0.040	0.039	0.032	0.027

10	0.040	0.002	0.001	0.068
11	0.040	0.016	0.013	0.013
12	0.040	0.025	0.036	0.029
13	0.040	0.054	0.032	0.017
14	0.040	0.035	0.004	0.038

5.6 ACDM: Results Analysis

As shown in the fig. 2, the threshold probability value (Pth) can be compared with observed probability values computed during testing phase. In the figure, threshold probability value is shown with a red line and the area that lies below this trend line are web-based attacks that are detected and values above the red line are the attacks that are not detected by the ACDM.

The FAR for ACDM can be calculated using the equation given below and results are shown in the Table 6.

$$FAR = \frac{Number\ of\ malicious\ query\ accepted}{Total\ number\ of\ queries}$$

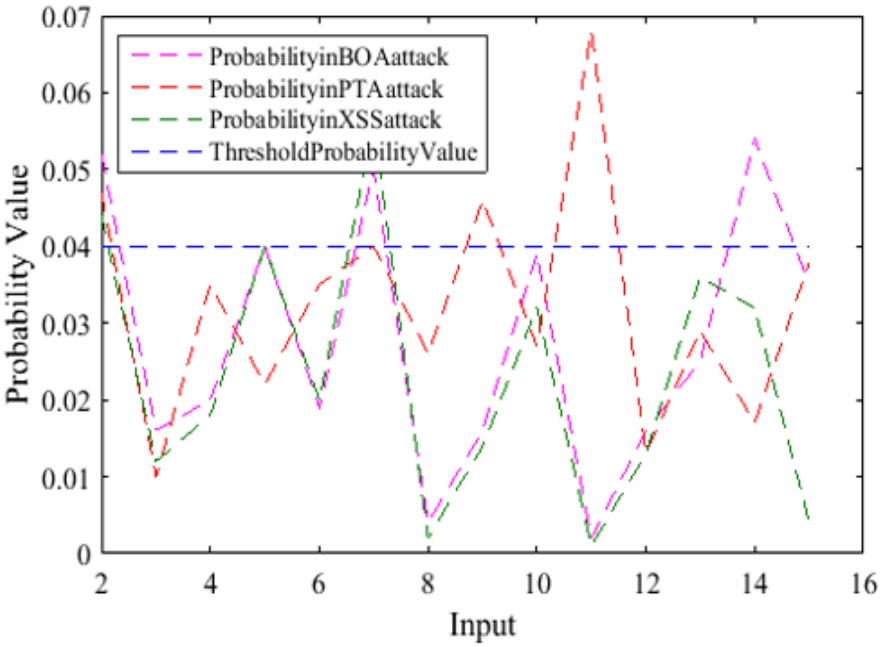

Fig. 2: Threshold probability values against attacks probability values

Table 6: FAR calculation in ACDM

Variables	**Values**
Total number of queries	42
Number of malicious queries accepted	11
FAR	26.1%

As FAR shown in the table for ACDM, it is clear that only 26.1% web-based attacks are not successfully detected the area above the red line in the figure. It means, we can say that the ACDM successfully detected 73.9% web attacks the area below the red line in the figure.

5.7Comparison between the ALM and ACDM

In this paper, we have analyzed the anomaly based web-based attack detection methods for three different attacks BOA, XSS and PTA. In fig. 3, the FAR for both the methods ALM and ACDM are shown for each of the three web attacks.

The FAR comparison for all the three attacks has shown in fig. 3. As shown in the figure, ALM is best suited for the detection of all three web attacks that are considered in this work. In case of BOA, ALM has successfully detected all the malicious queries whereas FAR for ACDM is 28%. For PTA, FAR is 16% for ALM and FAR is 28% for ACDM. It means ALM is better than ACDM to detect the malicious queries. Similarly in case of XSS attack, FAR is 12% for ALM and 21% for ACDM. Therefore, after analysis of all these web attacks, we can say that ALM is more efficient method than ACDM in detection of web based attacks.

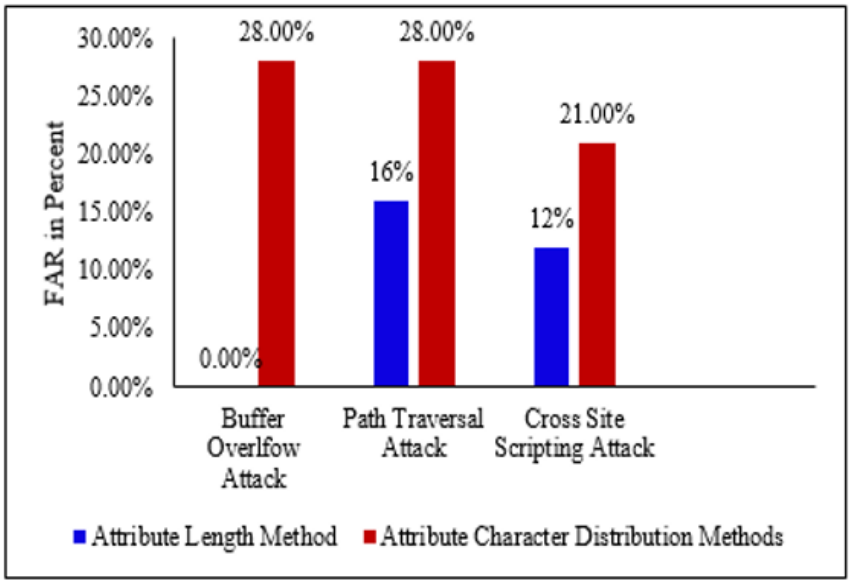

Fig. 3: ALM and ACDM FAR Comparison

5. CONCLUSION

In this paper we have analytically compared the two anomaly detection methods i.e. ALM and ACDM. We have done the mathematical analysis of both the methods on three different web attacks i.e. BOA, XSS and PTA. Also we have compared their False Accept Rate (FAR) results for both the methods. Experimental results shows that both the methods are capable of detecting web based attack very efficiently as in both the methods the FAR is less than 30% and further results analysis revels that ALM is more efficient method than ACDM in detection of web based attacks.

REFERENCES

[1] Christopher Kruegel , Giovanni Vigna , William Robertson, A multi-model approach to the detection of web-based attacks, Computer Networks: The International Journal of Computer and Telecommunications Networking, v.48 n.5, p.717-738, 5 August 2005.

[2] A. Singhal and S. Jajodia, "Data warehousing and data mining techniques for intrusion detection systems," Distributed and Parallel Databases, vol. 20, pp. 149-166,2006.

[3] H. Debar, M. Dacier, and A. Wespi, "Towards a taxonomy of intrusion-detection systems," presented at Computer Networks, 1999.

[4] Open Source Vulnerability Database (OSVDB), Accessed from http://osvdb.org, November 2011

[5] Common Vulnerabilities and Exposures (CVE), Accessed from http://cve.mitre.org, November 2011

[6] M. -L. Shyu, S. -C. Chen, K. Sarinnapakorn, L. Chang, (2003) A novel anomaly detection scheme based on principal component classifier, In Proceedings of the 3rd IEEE International Conference on Data Mining, pp. 172–179.

[7] C. Kruegel, G. Vigna, W. Robertson,(2005) A multi-model approach to the detection of web-based attacks, Computer Networks 48 (5) , pp 717–738

[8] C. Noble and D. Cook. (2003) Graph-based anomaly detection. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 631–636.

[9] Maxion RA, Tan KMC (2000) Benchmarking anomaly-based detection systems. In: International Conference on Dependable Systems and Networks. IEEE Computer Society Press, Los Alamitos, pp 623–630.

[10] Este´vez-Tapiador J. M. ,Garc?´a-Teodoro P. , D?´az-Verdejo J. E. (2005) Detection of web-based attacks through Markovian protocol parsing. In: Proc. ISCC05; pp. 457–62

[11] J. Gomez, D. Dasgupta,(2001) Evolving fuzzy classifiers for intrusion detection, in: Proceedings of IEEE Workshop on Information Assurance, United State Military Academy, West Point, NY, 2001, pp. 68–75.

[12] J. M. Estévez-Tapiador, P. García-Teodoro, J. E. DíazVerdejo (2004), "Measuring Normality in HTTP Traffic for Anomaly-Based Intrusion Detection", in. Computer Networks, 45(2), pp 145-193.

[13] Tejinder Singh Mehta , Sanjay Jamwal, (2015) Model To Prevent Websites From XSS Vulnerabilities, (IJCSIT) International Journal of Computer Science and Information Technologies, 6 (2) , pp 1059-1067

[14] OWASP Top 10 Application Security Risks, Accessed from http://www.owasp.org/index.php/Top_10_2010-Main, November 2011

[15] CERT. Advisory CA-2000-02: malicious HTML tags embedded in client web requests. Accessed from http://www.cert.org/advisories/CA-2000-02.html, 2000

[16] David Endler. The Evolution of Cross Site Scripting Attacks. Technical report, iDEFENSE Labs, 2002

[17] Scott Berinato, Software Vulnerability Disclosure: The Chilling Effect, 2007, Accessed from http://www.csoonline.com/article/221113/software-vulnerability-disclosure-the-chilling-effect

[18] Kuperman, B. A., Brodley, C. E., Ozdoganoglu, H., Vijaykumar, T. N., and Jalote, A. Detecting and prevention of stack buffer overflow attacks. Communications of the ACM 48, 11 (2005)

[19] OWASP, Path Traversal Attack, Accessed from https://www.owasp.org/index.php/Category:Path_Traversal_Attack.

[20] Patil, Mr Sachin S., Deepak Kapgate, and P. S. Prasad. "Effective Concept for Detection of Web Based Attacks Using ID3 Algorithm." (2014).

[21] Agrawal, Snigdha, et al. "Detection and Implementation of Web-based Attacks using Attribute Length Method." International Journal of Computer Applications 120.3 (2015).

[22] Boggs, Nathaniel, et al. "Synthetic Data Generation and Defense in Depth Measurement of Web Applications." Research in Attacks, Intrusions and Defenses. Springer International Publishing, 2014. 234-254.

[23] Massa, Daniel, and Raul Valverde. "A fraud detection system based on anomaly intrusion detection systems for e- commerce applications." Computer and Information Science 7.2 (2014): p117.

[24] Han, EiEi. "Detection of Web Application Attacks with Request Length Module and Regex Pattern Analysis." Genetic and Evolutionary Computing. Springer International Publishing, 2015. 157-165.

AUTHORS

Achin Jain, working as Assistant Professor in BharatiVidyapeeth College of Engineering, New Delhi (Affiliated to Guru Gobind Singh Indraprastha University, Delhi) in the Information Technology Department. He received his M.Tech(Computer Science and Technology) &B.Tech(Information Technology) from the Guru Gobind Singh Indraprastha University, Delhi. He has rich experience teaching B.Tech students and has published more than 6 Research Papers in International Journals and Conferences. His area of interest includes Web Usage Mining, Web Attacks.

Vanita Jain received her Ph.D. degree in Electrical Engineering from V.J.T.I.,Mumbai, India , M.Tech. degree in Control Systems from National Institute of Technology, Kurukshetra and B.E. degreen from Punjab Engineering College, Chandigarh, India. From 1998 to 1999, she worked at National Institute of Technology, Kurukshetra as a lecturer. She joined ThadomalShahaniEngineeringCollege, Bandra, Mumbai and served in the Department of Electronics & Telecommunication Engineering for 15 years. Currently, she is Dean (Academics) at BharatiVidyapeeth's College of Engineering, New Delhi. She is having more than 24 years of teaching experience. Her field of interest includes Soft Computing, Control, Optimization Techniques and System Engineering and is actively involved in the research in these areas.