International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-4, April 2019

Detection and Prevention Approach to SQLi and Phishing Attack using Machine Learning

J.Jagadessan, Akshat Shrivastava, Arman Ansari, Laxmi Kanta Kar, Mukul Kumar

Abstract*: Web application attacks are increasing and exploiting the security of users. The flow of our paper goes with the discussion of cyber security attacks to the machine learning algorithms to detect and prevent these types of attacks. This paper also uses an open source Web Application Firewall- Mod Security which is an internet technology helps preventing the attacks. We have discussed the approach of WAF with the algorithms of machine learning to efficiently detect the attacks and secure the user. Machine learning learns the attack from previous attacks according to the previous results and block or bypass the Web Applications Firewall. The paper focuses on SQL Injections and Phishing vulnerabilities and prevent attackers to easily deceive them.

Index Terms*: - Web Application Firewalls (WAF), Machine Learning, Mod Security, Cyber Security attacks, SQL Injections, Phishing.

I. INTRODUCTION

A Web Application is an application software that uses online web browser to perform different task over the internet. Millions of businesses and common people use the internet for different purpose like cost-effective communication channels, storing data, making transactions, accessing social networking sites. These Web Applications are easily exploited and attacked by hackers and crackers. However, these attacks can be prevented by using the concepts of Web Application Firewall and Machine Learning. WAF enables the user to access the real time Web Application monitoring and its access control. Firewall is a network security system that can control and monitor the traffic on the network based on some rules that are predefined for the security purpose. In actual the firewall establishes a barrier between a trusted network and an untrusted network which can be an internet.

Manuscript published on 30 April 2019.
***** Correspondence Author (s)
Dr. J. Jagadessan*, Head of Department, Department of CSE, SRM Institute of Science and Technology, Ramapuram, Chennai
Akshat Shrivastava, Student, Department of CSE, SRM Institute of Science and Technology, Ramapuram, Chennai
Md Arman Ansari, Student, Department of CSE, SRM Institute of Science and Technology Ramapuram, Chennai
Laxmi Kanta Kar, Student, Department of CSE, SRM Institute of Science and Technology, Ramapuram, Chennai
Mukul Kumar, Student, Department of CSE, SRM Institute of Science and Technology, Ramapuram, Chennai

There is category of firewall- Network firewall and host-based firewall. Network based firewall works on LANs, WANs and intranets. These can be a software application running on a hardware or a hardware firewall itself. Firewall can also possess others functions such as VPN servers. Host based firewall are positioned on the network node for controlling network traffic to and from the machine. These firewalls are divided into types which are network layers or packet filters, Applications layers, Proxys and network address translations. Packet filters are positioned on low level of TCP/IP Protocols, basically disapprove the packet to pass through
firewall unless they are matched with predefined rule set. A Proxy server also act as firewall by responding to packets at the receiving end in the form of an applications while blocking others packets. Network address translations is a function of firewall and protect the host that have addresses in the private address range. The last type of firewall which is Applications layer firewall is a type of firewall that defines Web Applications firewall which work in the application level of TCP/IP stack that obstruct all packet travelling in and out of an application. Mod Security is an open source tool kit for real time monitoring of web application and its access control. The four guiding principles on which these tool kits are based are flexibility, passiveness, predictability and quality over quantity. It is cross platform WAF module which enables web application defenders to produce the visibility into the traffic of HTTP and provides rules language and API for implementing other security protection. As discussed, these Web Application Firewall bypass the cyber security attacks that may try to cause harm. Most common types of attacks that are seen now-a-days are- Malware, XSS, DoS, Session hijacking, Information Reuse, Phishing and SQL Injections. Malware are harmful software such as ransomware and viruses. Malware can harm the computers and take control of the machine to silently monitor the action and keys strokes that can release the confidential data from the user. Cross-Site Scripting can lead to attack and target a website user with a loop hole and targets its data such as credentials and financial data by injecting harmful code into a website. In Denial of Service a flood is created on a website to increase the traffic more than it was built to handle and make the website content unavailable to the users accessing it.

Session Hijacking will be seizing of session through by catching the session id and propose as a PC making a demand which enable them to login and obtain entrance as an unapproved client. Most people reuse the same credentials which has been used previously. The Credential Reuse is the easiest type of attack in which the attackers have a collection of password and username from a breached website which is easily available on black market website. These credentials can be the same that a user is using currently and the attacker can gain the access to his/her e-mail, bank account, social networking, etc. The Phishing attack pretends to a trustworthy website or an email where the users are asked to enter some personal information's or credentials. In these attacks the attacker mimics a trustworthy website which has a legitimate looking and make a trap to capture the details. This paper focus on these attacks which can be harmful for the users or can bypass any firewall. SQL Injections attacks target the database server using malicious code and gain critical data stored in the database. These attacks are done by exploiting SQL vulnerabilities allowing the SQL server to run the harmful code.

Attackers come out with different proposed attacks with new loop holes and vulnerabilities which a WAF cannot bypass or prevent. In this case Machine Learning comes to play which automatically increment the loop holes and train the WAF to bypass or prevent it.

The Machine Learning methods discussed in this paper for Phishing attacks are- Model selection for support vector machine (SVM), biased support vector machine (BSVM), neural networks and K-means [7]. Whereas, the machine learning algorithm for detecting SQLi attacks are Random forest (Random tree) [6] with the use of classifier algorithm which are discussed in Section 6.2.

A. SQLi Attack

Web technologies are used in everyday business around the world. The web-based apps use the Relational Database which are operated by the Structured Query Language (SQL).SQL is a particular type of query driven language which is used to interact with the database and perform the operations. SQL injection is a technique which is used to exploit the websites by using certain characters or patterns through the queries. Generation of query from user side can be malicious. During the query execution, the web apps are vulnerable as they can take input from any external source which can be malicious for such type of intrusion the developer has to write efficient codes. Many user inputs can be exploit because of developer's illiterateness.

Some of the basic SQLi types includes Tautology attack, union-based attack, etc. In the tautology attack, the SQL query modification is processed in such that the result is always true. In the union-based attack, the UNION statement is combined with any sql query in a way that either of the queries acts maliciously.

Unauthorized access is a type of SQLIA in which the attacker uses tautology class for SQL injection by which it modifies the query and the attacker can use any password to login into the database. The WHERE statement is used to extract particular entities. For e.g.:

```
Select * From tbl_name Where id='abc01' AND
Password = 'xyz';
```

This statement is used to fetch data from a table named "tbl_name" where id and password matches from the database entries and data can be fetched from the back-end where if either of the Where statement clause is true. The SQL Injection results in a new query which is always TRUE. This changes the WHERE clause structure by which the attacker can access all the data. This can be shown by using the syntax given below

```
'OR 1=1;
```

The resultant query due to above mentioned syntax for sql injection is given below:

```
Select * From tbl_name Where id= ' OR 1 =1' -- ' AND
password='abc001';
```

After the processing of above query the attacker can access all the data as the WHERE statements result always TRUE as it always has condition 1=1.

After the successful modification the data is extracted from the table even if the given password and id is incorrect whereas the line after the comment statement are excluded.

B. Phishing Attack

Phishing is one of the cyber-security attacks where an attacker will obtain the personal information of the user which can be related to the accounts login details by sending a spam email or any other channel. The user information becomes vulnerable and are exploited with software which are sent by any organization which contains malicious code.

Phishing is one of the most popular technique because it is easy for the attacker to trick a user by sending any spam or malicious URL which seems similar to a known website. It is easier than trying to break through a computer's defence systems to exploit any user. The malicious links are developed in such a way that they represent a true organisation with its fake logo and other true contents. The users click the malicious link unwittingly and are exploited by the attackers easily.

The phishing attack is focused mainly on the URL which have some protocols which results to access the page information. The user's webpage is identified by the domain name and launched into the web server. The Registrar is used to constrain the portion of domain name.

A Host name has two parts: subdomain name and domain name. An attacker can set the value for any subdomain portion. The phisher can change the file components of the URL. The phisher can create a URL and have total control over it. The phisher uses the unique parts of the domain by which the security mechanism failed to detect the attack. The attacker has to choose a unique domain names because it's must look convincing the users.

.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

II. LITERATURE SURVEY AND RELATED WORK

In this section we have discussed different reference paper and their review which helped and motivated us in this paper.

A. Critical Analysis on Web Application Firewall Solutions [1]

The web application is being exploited very easily these days. Different solutions are discovered to protect the web applications from these various harmful attacks. This paper provides an overall analysis to the different solutions to these web application attacks. It also contains a report on the analysis of the different solution presented there.

B. A Study on Web Application Security and Detecting Security Vulnerabilities [5]

Web application securities deals with the protection of the web applications from the different attacks. The web applications have become vulnerable to various wider range of attacks. These attackers exploit the system and extract the content which are sensitive. The entire world depends on the web applications for the ongoing business transactions. This paper consists of the different solutions to the attacks which prevents the sensitive data from being exploited. It also provides full description of various attack mechanism and the defense system employed towards it.

C. Web Application Security Tools Analysis [2]

Web application security has been the most demanding task in the present scenario. The demand of web security has grown rapidly as the attackers have figure out various way to exploit the web applications. This paper provides the analysis report of the different security tools employed to prevent the web applications from being exploited. The tools provide the different mechanism and features which can be used to provide web application security.

D. Behind an Application Firewall, Are We Safe From SQL Injection Attacks? [3]

The web application firewall is a mechanism which prevents the web applications from the various vulnerabilities and attacks. This paper provides a study report on the firewalls. It describes the various updating and alteration required in the firewall to keep them up to date. The approach in this paper is to prevent the web applications from the SQLi attacks and focus on a machine learning algorithm. The various machine learning test approach has been used to detect the holes in the SQLi attack and learn from the pattern. The goal is to prepare the firewall in such a way that it can keep on updating itself on the attack pattern. Mod Security is the tool used in this paper to showcase the machine learning techniques employed.

E. *Automatically Repairing Web Application Firewalls Based on Successful SQL Injection Attacks* [4]

The challenges faced by the web app developers these days is the security of the web applications and their protection. The automated testing techniques helps in reducing the cost and find out the different. The solution of the vulnerabilities is searched and found out to fix the issue after it has completely detected the attack. The objective of this paper is to optimize the problem with different machine learning techniques.

Explanation

Since, the techniques used in the preliminary work for these critical security risk that are SQL Injection attack and phishing attack make use of the techniques which has some limitations that can impact the practical application and their vulnerability detection capability. The techniques that are used in the reference paper for detecting and preventing SQLi attacks are white box testing method and some tools for analysis, which requires credential permission to reach to the source code which may or may not be in the reach with some other specific party component, and are dependent to any source programming language. Whereas, Model based testing method requires modelling that expresses the security protocols or the implementation of web application firewall and also the application used under testing, which are very difficult for construction. The black box texting method doesn't require the strategies that are used with white box testing or model-based techniques but are less effective in detecting the loop holes of SQL Injection [6].

The phishing attacks in this model internet word has spread rapidly. These attacks can lead to many financial and similar losses. It is very difficult to trace the hacker. Thus, the first solution for preventing these types of attacks start from the awareness from the user which may not be a successful method. The non-technical method which is a legal solution to the problem. In many countries which requires the task to trace the hackers which is not 100% successful approach. In technical method there can be two approach the first is Blacklist technique which is a database of pre-established phishing techniques or websites. Thus, it doesn't deal with each and every phishing website because there can be establishment of new fake website every day. The second approach is a fuzzy role approach which includes gathering of features. These techniques have not seemed to be effective in preventing and detecting the phishing attack [7].

III. PROPOSED SYSTEM

Looking into the preliminary work in the detection of these cyber-attacks, the efficiency of the prevention and detection is not satisfactory. Therefore, in this paper, we have proposed the concept of machine learning to train the system and detect these attacks efficiently and securely. For phishing attacks there can be spam emails or phishing emails which can be detected using various machine learning techniques and clustering methods on the phishing data sets for this support vector machine, composed neural networks, organizing maps and k-means technique are used. Whereas the phishing domain detection implementing machine learning have a different approach. These machine learning techniques can be grouped into the following features, those are URL-based, Domain based, page based and content based. The detection is a classification problem. Therefore, the data sets for phishing is done from an open source, phish tank which is a commonly used data source in academic studies. The algorithm used in detecting these attacks in decision tree algorithm, which is a simple and powerful as described in Section 6.1.2.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

In SQLi attacks, different attacks strategies are combined to get some better result with the help of machine learning algorithms in which context free grammar is used to define the input space after which a grammar based random attack generation is used as a simple attack generation strategy to sample the input. Now, for a machine learning guided attack generation to guide the test generation in the input space, ML-Driven approaches are used which are attack decomposition, training set preparations and the path condition in which the procedure used is Random Forest consisting of multiple random trees. To create different offspring patterns to protect the user and bypass or block the attack. This ML-Driven approach can be enhanced to balance exploration and exploitation by using ML-Driven B and ML-Driven D where ML-Driven D where ML – Driven D leads to select fewer tests for mutations and generating more mutants, the balance is set to high in this approach. For ML-Driven B the balance is specified to a lower value that means collecting other tests for mutations and generating fewer mutants. However, these approaches can be combined to increase the efficiency. The subject application used for the evaluation is an open source WAF ModSecurity. Also, some selected temper scripts of SqlMap form the source sqlmap.org is used as an experiment.

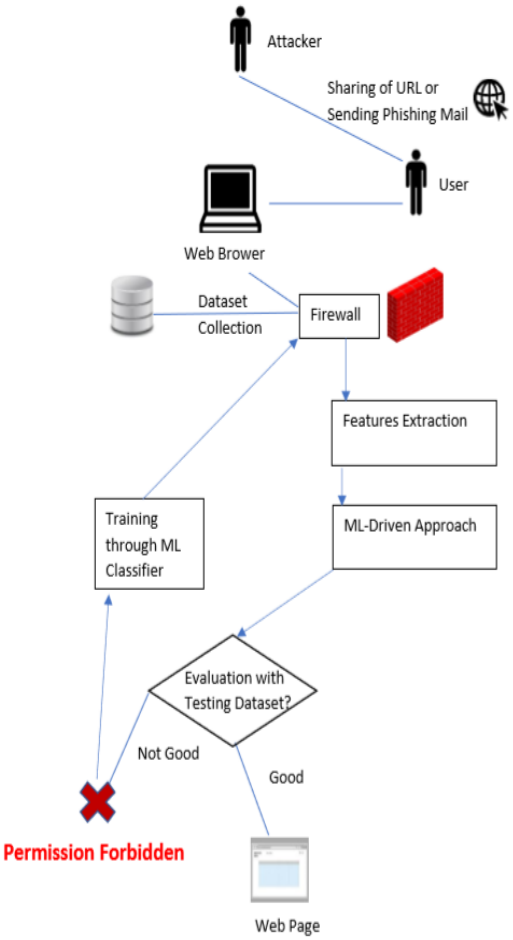

Fig. 1: Preventing SQLi Attack

IV. SYSTEM ACHITECTURE

A. SQL Injection Attack Prevention

The architecture explains that when an attacker tries to attack the server by SQLi through a web browser entering some malicious code and authorized with the access into the database server. The firewall here allows the attacker to access the database server by injecting the malicious SQL queries and gain an unauthorised access to the database server.This firewall is can be Open Source which are stock or in-built with the antivirus software these days.

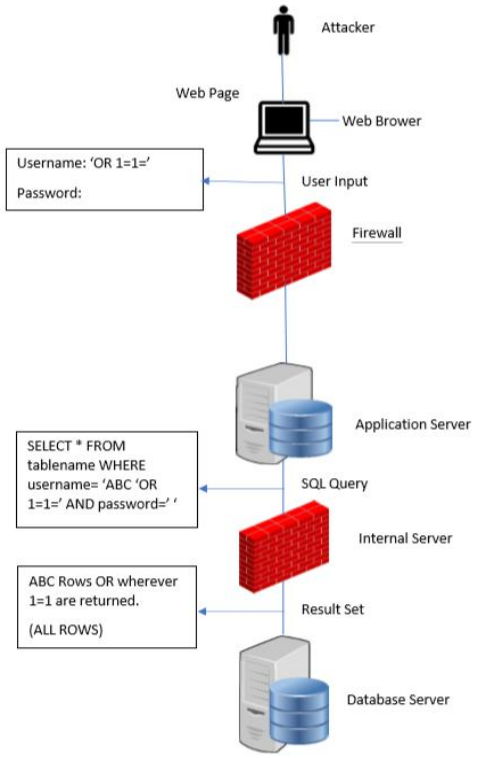

Fig. 2: Successful SQL Injection

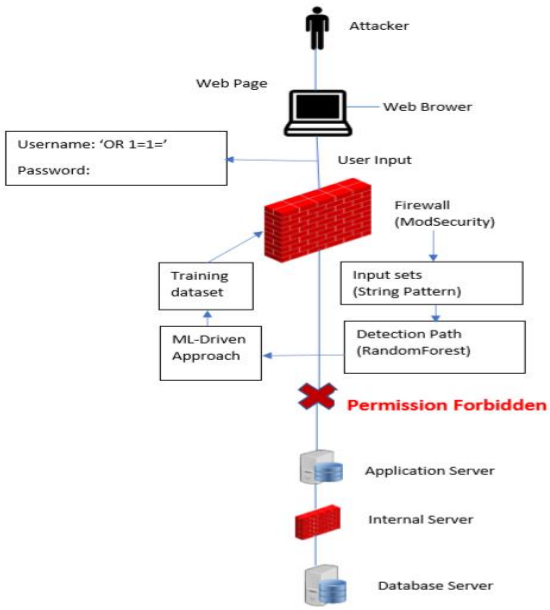

Fig. 3: Prevention of Phishing Attack

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

The architecture above clearly explains that when using a web firewall which is ModSecurity in this case denies the permission to access the database server when an attacker tries to inject some malicious SQL queries. The firewall train itself by using the Machine Learning approach and input.

B. Phishing Attack Prevention

The architecture for preventing the Phishing Attack is a Machine Learning implemented approach where first an attacker shares the URL or send the mail to the user.

The user then tries to open the URL or E-mail thinking it as a intimidate process, but the Firewall or Security System denies the permission to do so and train itself through the machine learning approach. Now if the Firewall detects the mail or the webpage as safe, it allows the user to access it. The Firewall in the starting train itself from a preliminary dataset acquired from a source.

V. DATASET USED

A. *Phishing URLs*

The dataset used for the training the firewall to detect and prevent the phishing URL is downloaded from the source PhishTank.

phish_id	url	phish_detail_url	submissior	verified
5949481	https://rnyuthewallet.xyz/	http://www.phishtank.cc	2019-02-2	yes
5949473	https://sada-e-abuzar.tv/wp-admin/user/ck/sign	http://www.phishtank.cc	2019-02-2	yes
5949472	https://vendedormundial.com/scib/Absa%20Onl	http://www.phishtank.cc	2019-02-2	yes
5949471	https://www.rentalwebsiteinaweek.com/scib/Al	http://www.phishtank.cc	2019-02-2	yes
5949470	http://peplz.pl/scib/Absa%20Online%20-%20PRC	http://www.phishtank.cc	2019-02-2	yes
5949467	http://www.web-mydocomo.com/	http://www.phishtank.cc	2019-02-2	yes
5949466	http://web-mydocomo.com/	http://www.phishtank.cc	2019-02-2	yes
5949465	http://thmees.myvnc.com/confirm/PayPai/servi	http://www.phishtank.cc	2019-02-2	yes
5949464	http://thmees.myvnc.com/confirm/PayPai/servi	http://www.phishtank.cc	2019-02-2	yes
5949463	http://drohobych-prolife.org.ua/www/login/hon	http://www.phishtank.cc	2019-02-2	yes

Fig. 4: URL based Phishing Dataset

B. *Phishing E-mails*

The dataset used for detecting phishing emails is downloaded from SpamAssassin and PhishingCorpus and is as follows


fradulent_emails - Notepad

File Edit Format View Help

From r Wed Oct 30 21:41:56 2002
Return-Path: <james_ngola2002@maktoob.com>
X-Sieve: cmu-sieve 2.0
Return-Path: <james_ngola2002@maktoob.com>
Message-Id: <200210310241.g9V2fNm002828@cs.CU>
From: "MR. JAMES NGOLA." <james_ngola2002@maktoob.com>
Reply-To: james_ngola2002@maktoob.com
To: webmaster@aclweb.org
Date: Thu, 31 Oct 2002 02:38:20 +0000
Subject: URGENT BUSINESS ASSISTANCE AND PARTNERSHIP
X-Mailer: Microsoft Outlook Express 5.00.2910.0000 DM
MIME-Version: 1.0
Content-Type: text/plain; charset="us-ascii"
Content-Transfer-Encoding: 8bit
X-MIME-Autoconverted: from quoted-printable to 8bit by sideshowmel.si.UM id g9V2fGW24311
Status: O

FROM:MR. JAMES NGOLA.
CONFIDENTIAL TEL: 233-27-587980.
E-MAIL: (james_ngola2002@maktoob.com).

URGENT BUSINESS ASSISTANCE AND PARTNERSHIP.

DEAR FRIEND,

I AM (DR.) JAMES NGOLA, THE PERSONAL ASSISTANCE TO THE LATE CONGOLESE (PRESIDENT LAURENT KABILA) WHO WAS ASSASSINATED BY HIS BODY GUARD ON 16TH JAN. 2001.

THE INCIDENT OCCURRED IN OUR PRESENCE WHILE WE WERE HOLDING MEETING WITH HIS EXCELLENCY OVER THE FINANCIAL RETURNS FROM THE DIAMOND SALES IN THE AREAS CONTR

MY PURPOSE OF WRITING YOU THIS LETTER IS TO SOLICIT FOR YOUR ASSISTANCE AS TO BE A COVER TO THE FUND AND ALSO COLLABORATION IN MOVING THE SAID FUND INTO YOU

YOU WERE INTRODUCED TO ME BY A RELIABLE FRIEND OF MINE WHO IS A TRAVELLER,AND ALSO A MEMBER OF CHAMBER OF COMMERCE AS A RELIABLE AND TRUSTWORTHY PERSON WHOM

THE (USD$25M) WAS PART OF A PROCEEDS FROM DIAMOND TRADE MEANT FOR THE LATE PRESIDENT LAURENT KABILA WHICH WAS DELIVERED THROUGH ZIMBABWE IN DIPLOMATIC BOXES

AS A MATTER OF FACT, WHAT I URGENTLY NEEDED FROM YOU IS YOUR ASSISTANCE IN MOVING THIS MONEY INTO YOUR ACCOUNT IN YOUR COUNTRY FOR INVESTMENT WITHOUT RAISIN

THE REMAINING 5% WILL BE USED TO OFFSET ANY COST INCURRED IN THE CAUSE OF MOVING THE MONEY TO YOUR ACCOUNT. IF THE PROPOSAL IS ACCEPTABLE TO YOU PLEASE CONT

FINALLY, IT IS IMPORTANT ALSO THAT I LET YOU UNDERSTAND THAT THERE IS NO RISK INVOLVED WHATSOEVER AS THE MONEY HAD NO RECORD IN KINSHASA FOR IT WAS MEANT FO

LOOKING FORWARD TO YOUR URGENT RESPONSE.

YOUR SINCERELY,

MR. JAMES NGOLA.

Fig. 5: Mail based Phishing Dataset

C. *SQL Injection*

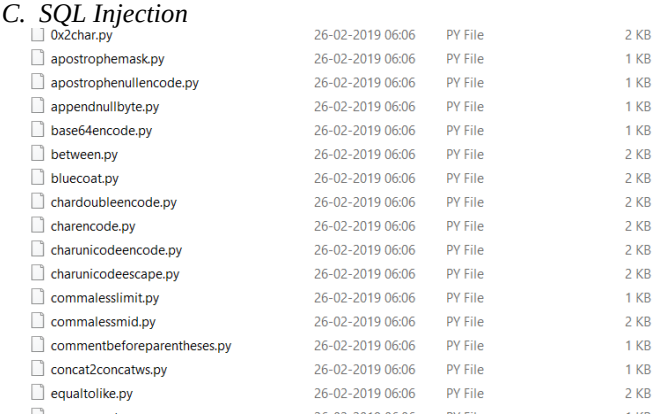
0x2char.py 26-02-2019 06:06 PY File 2 KB
apostrophemask.py 26-02-2019 06:06 PY File 1 KB
apostrophenullencode.py 26-02-2019 06:06 PY File 1 KB
appendnullbyte.py 26-02-2019 06:06 PY File 1 KB
base64encode.py 26-02-2019 06:06 PY File 1 KB
between.py 26-02-2019 06:06 PY File 2 KB
bluecoat.py 26-02-2019 06:06 PY File 2 KB
chardoubleencode.py 26-02-2019 06:06 PY File 2 KB
charencode.py 26-02-2019 06:06 PY File 2 KB
charunicodeencode.py 26-02-2019 06:06 PY File 2 KB
charunicodeescape.py 26-02-2019 06:06 PY File 2 KB
commalesslimit.py 26-02-2019 06:06 PY File 1 KB
commalessmid.py 26-02-2019 06:06 PY File 2 KB
commentbeforeparentheses.py 26-02-2019 06:06 PY File 1 KB
concat2concatws.py 26-02-2019 06:06 PY File 1 KB
equaltolike.py 26-02-2019 06:06 PY File 2 KB

Fig. 6: SQLi Tamper Script

The tamper script used as a dataset for training the web application firewall against SQLi attacks has source from sqlmap.org website. Some of the python codes are as follows

VIII. APPROACHES

The proposed approach for detecting and preventing the SQLi attacks and phishing attack are explained in the section. This paper uses different machine learning methods which uses the dataset for updating and incrementing the database.

1. *Machine learning approach for phishing attack*

As discussed in Section 1 that phishing can be done by either a spam e-mail or duplicates illegal website.

1.1 *Phishing E-mails*

Therefore, for the phishing e-mails and legitimate e-mails, open source website can be used which divide the entire set into two parts for testing as well as for training purpose. To pass the phishing and legitimate e-mails the python and Java script is used in which different features as described below:

1.1.1 *Features used*

i. *HTML image*- Most of the used HTML formatted e-mails can be used for phishing attack because the text e-mail doesn't provide the scale of tricks. And hyperlinks or the URLs which are clickable are done in HTML formatted e-mail. Thus, this feature can be flagged.

ii. *IP based domain*- When an attacker uses the IP address instead of domain, it makes the user hard to process where they are being directed to whereas a true or legitimate have a domain name.

iii. *Age of domain name*- Fraudsters usually makes use of domains name for a short period of time to avoid being caught. Thus, this feature can be used to flag as phishing process based with the time of registration of domain.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

iv. *Number of Sub domains*- The sub domains name is used by attackers to make the URL look legitimate. Thus, the structure of sub domain (large number of dot) can be a feature to flag image as phishing.

v. *Presence of JAVA Script*- As Java Script enables to make changes on the client side, these are usually used in phishing image.

vi. *Authenticity*- With Social-engineering with different attackers used keywords, website based, images source to make the e-mail look true and authentic. Thus, this binary feature can be used to flag e-mails as phishing e-mails.

vii. *Form tag*- HTML forms in e-mail may be used to collect the information in which the attacker actually sells the information on submitted by the user to his/her pre-own website replicating some legitimate website.

1.1.2 *Machine Learning Approach*

The different machine learning techniques and clustering techniques or methods are as follows:

i. *Model of support vector Machines*- In probable learning task which can be classification, a modeling and parameters prediction approach should be used to get great performance of learning machines. Within the SPVM approach, the parameter that used are- the penalty term F' which usually determine the tradeoff between the number of training example and the complexity of decision function, the mapping function **C** and the Kernel function as:

F(xi,xj)=C(xi). C(xj)

candidateSV = { closest pair from opposite classes }
while there are violating points **do**
  Find a violator
  candidateSV = candidateSV ⋃ violator
  if any $\alpha_p < 0$ due to addition of c to S **then**
    candidateSV = candidateSV \ p
    repeat till all such points are pruned
  end if
end while

Fig. 7: SVM Algorithm

ii. *Biased support vector machine*- BSVM is one of the decomposition techniques to support vector machines for a larger number of classification problem. This decomposition method can reduce and solve a bound constraints SVM formulations. BSVM uses set selection which can lead to fast convergences for any difficult input and a bounded SVM formulation which allows it to quickly identify support vector.

iii. *Neural Networks*- Artificial Neural network is an accumulation of preparing element which are interconnected and change contributions to set off wanted or desired output. The result of this transformation is determined through the characteristics of the element and the weights associated with interconnection. A neural system ordinarily leads an investigation of the information and produces an expected likelihood that it matches with the data it has been prepared to distinguish. The neural systems pick up the experience at first via preparing a framework with info and yield of the issue.

iv. *K-means*- This is an unsupervised non-hierarchical bunching technique which endeavors to improve the estimation of the mean of each group and renames each group with closest mean. The methodology utilizes iterative strategy which combines to one of the various neighborhoods focuses. This iterative methodology is sensitive to beginning position.

1.2 *Phishing Websites*

The Phishing domain detections use certain types of features which can be blocked to avoid attacks. For example, any attacker can register a legitimate looking website to fool the user and gain some information.

1.2.1 *Features used*

Different type of feature which are used in Machine Learning:

i. URL based feature- URL is one thing to examine a site to choose whether it is destructive or not. URLs can have some unmistakable focuses. Cases of URL based highlights can be digit counting, complete length, and typo squatted or not, an authentic brand name or not, number of sub-areas and is TLD-Top Level Domain normally utilized?.

ii. Domain based feature- This recognizes domain name of the phishing site. Aloof inquiries identified with the name of domain which can be characterize as phishing or not, give some helpful data. This highlights incorporate the domain name or its IP address is blacklisted in some database or not, the days gone since the domain was enrolled and Is the registrant name covered up?.

iii. Page based feature- This feature are data about the pages which have calculated reputation ranking services. This decides if the website is reliable or not. Some of page-based features are global page rank, country page rank, estimated number of visits, web traffic, and similar website.

iv. Content based features- This feature requires an active scan to detect a phishing domain. Page contents are processed in this feature. The contents can be page titles, hidden tax, meta description and images. This provides the information to analyse whether to login into the website, the information about audience profile, etc.

1.2.2 Data Pre-processing

A URL comprise of some helpful and pointless words and exceptional characters which separate the most imperative parts of the URL. For instance, a speck imprint can be utilized to isolate SLD/TLD in like manner a way has a sub organizer are isolated utilizing forward slash /. In this way in information pre-preparing each word is removed in the location and they are investigated in the execution and the comparability or the words with focused sites and haphazardly made words are recognized in the module. Main aims of this process include detecting the word similar to word names and detecting the keyword in the URL.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

1.2.3 Word Decomposer Module (WDM)

WDM ordinarily dissect the URL and procedure it into isolated words on the grounds that an attacker can include numeric qualities for picking the values progressively complex the modules remove the digits first. After that the rest of the string is broke down and checked for its validating in the lexicon. In the event, the word is the lexicon word which is added to a list else it is isolated into substrings to establish a nearby word. The execution stream of the word decomposer is recursive. These are arranged by the length of URL from longest to most limited.

1.2.4 Maliciousness and analysis module

To detect whether an URL is fraudulent or not, this module is used to analyse the typo squatting. The module gets words as input and analyses accordingly as in the previous module a database prepared. It compares the database with the new incoming URL.

The detection process is a classification problem that means it requires labelled data which has sample as domain of phishing websites and true domains in the training sector. The dataset that are created for phishing have source from the PhishTank, which is commonly used in academic studies. Comprising+ the domains of legitimate website have the sources from site reputation services which analyse and rank available website. The Machine Learning algorithm has its own working mechanism. The algorithm used is decision tree algorithm which is simple and powerful. The first task is the need of labelled instances to create detection mechanism. This is done by done by two classes- Phishing and legitimate. The decision tree can be considered as improved version of nested if-else where each feature is checked one by one. Generating a tree is the super structure of the mechanism. The length of the tree is checked when an example arrives and other features are checked as per result. After the sampling the class of the sample is clear.

Decision tree uses information gain measure to indicate how a feature separate the example of training according to the target classification, which decides the root element. The method is known as the information gain and the mathematical equation is

$$Gain(S,A) = \underbrace{Entropy(S)}_{\text{original entropy of S}} - \underbrace{\sum_{v \in values(A)} \frac{|S_v|}{|S|} \cdot Entropy(S_v)}_{\text{relative entropy of S}}$$

High gain score indicates that the feature has high distinguishing ability because of which the feature is selected as root.

```
Tree-Learning (TR, Target, Attr)
    TR: training examples
    Target: target attribute
    Attr: set of descriptive attributes
{
    Create a Root node for the tree.
    If TR have the same target attribute value v_i,
      Then Return the single-node tree, i.e. Root, with target attribute = v_i
    If Attr = empty (i.e. there is no descriptive attributes available),
      Then Return  the single-node tree, i.e. Root, with most common value of Target in TR
    Otherwise
    {
      Select attribute A from Attr that best classify TR based on an entropy-based measure
      Set A the attribute for Root
      For each legal value of A, v_i, do
      {
        Add a branch below Root, corresponding to A = v_i
        Let TR_vi be the subset of TR that have A = v_i
        If TR_vi is empty,
          Then add a leaf node below the branch with target value =  most common value of
                Target in TR
          Else below the branch, add the subtree learned by
                Tree-Learning(TR_vi, Target, Attr-{A})
      }
    }
    Return (Root)
}
```

Fig. 8: Decision Tree Algorithm

In the phase of training the dataset is divided into parts by comparing the values of features. The algorithm calculates the information for each feature and select the one with maximum gain score. When the tree reaches a level the process of training is completed. The training set includes a large variety of sample from a large variety of dataset. The formulation methods given below are used to analyse the tested machines and learning approaches:

$$Precision = \frac{TP}{TP + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$F - Measure = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

Here, TN- true negative, TP- true positive, FP- false positive, FN- False Negative of classification algorithm.

2. Machine Learning approach for SQL Injection attacks

The machine learning approach for SQLi attack make use of Random forest and algorithm ML-Driven B and D which are discussed in this section.

2.1 Generation of inputs set

SQLi attacks are small queries which intend to target the database server. These attack datasets used in this paper have an open source SQLMap (sqlmap.org). We define a CFG for analysing the different attacks. The 3 categories of attacks in the grammar are Boolean Attack in which a SQL statement is either true or false, Union Attack which appends the results of multiple select statements, and Piggy-bag Attacks in which the statements are separated by the use of semicolon. The grammar for SQLi attacks in Backus normal form is easily accessible in GitHub. The grammar covers different context in which the attacks can be inserted into. This grammar is acting an input to the approach.

2.2 Random Attack Generation

According to the grammar this is a recursive straight forward method which is connected until no terminals are left to choose production rules. The output attack is produced by annexing the terminal images showed up in the instance of grammar. The production rule is chosen with likelihood extent to the quantity of unmistakable generation rules. The attack is then produced by re-applying the procedures on numerous occasions until a maximum number of tests is reached.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

2.3 Attack Generation Guided by Machine Learning

The techniques discussed in Section 2.2 become inefficient with large proportion of attacks. The Machine Learning approach first employ a random test generation explained in Section 2.2 to generate initial
training set. These tests are either bypassed or blocked by the WAF.
In the approach of search a derivation tree or parse tree is represented to detect the type of attack than a training set preparation is followed to label the test. After this a Path Condition or simply a Decision Tree is used to divide the test into slices and transform them into label dataset. The algorithms used for computing path conditions is Random Forest, which is a combination of multiple Random tress where different offspring or new test is induced.

Algorithm 2: ML-driven SQLi Attack Generation.

```
1:  procedure MLDRIVENGEN(initTests, outputTests)
2:      execute(initTests)
3:      P ← initTests
4:      archive ← UPDATEARCHIVE(P)
5:      // learn the initial classifier
6:      trainData ← transform(archive)
7:      DT ← learnClassifier(trainData)
8:      rankTests(P, DT)
9:      while not-done do
10:         O ← OFFSPRINGSGEN(P, λ, MAX_M)
11:         execute(O)
12:         archive ← UPDATEARCHIVE(O)
13:         // re-training the classifier
14:         trainData ← transform(archive)
15:         DT ← learnClassifier(trainData)
16:         // new population
17:         P ← SELECT(P ∪ O)
18:     end while
19:     outputTests ← filterBypassingTests(archive)
20:     return outputTests
21: end procedure
```

Fig. 9: ML-Driven Algorithm

This ML-Driven approach can be enhanced to balance exploration and exploitation by using ML-Driven B and ML-Driven D where ML-Driven D leads to select fewer tests for mutations and generating more mutants, the balance is set to high in this approach. For ML-Driven B the balance is set to a lower value that means it selects more tests for mutations and generating fewer mutants. However, these approaches can be combined to increase the efficiency

VII. EVALUATIONAL RESULTS

A. SQLi Attack

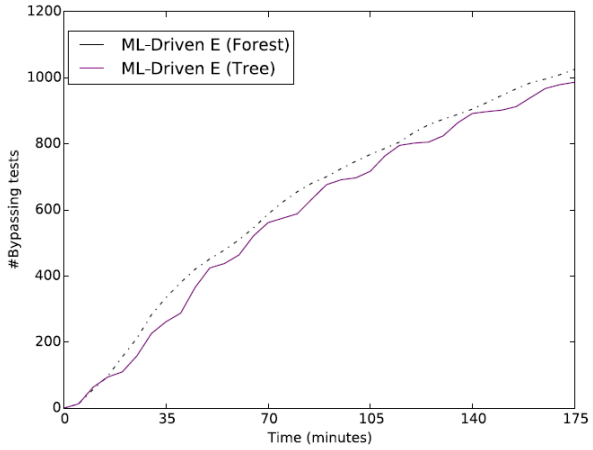

Fig. 10: Test results using ModSecurity

Using the results of WAF ModSecurity, the Random forest approach bypasses around 380 tests in 35 minutes whereas Random tree does it around 250 in 35 minutes.
This strategy is evaluated with an Open Source WAF ModSecurity which deployed with Apache HTTP Server which is used to secure a Web App.

B. Phishing – URL Attack

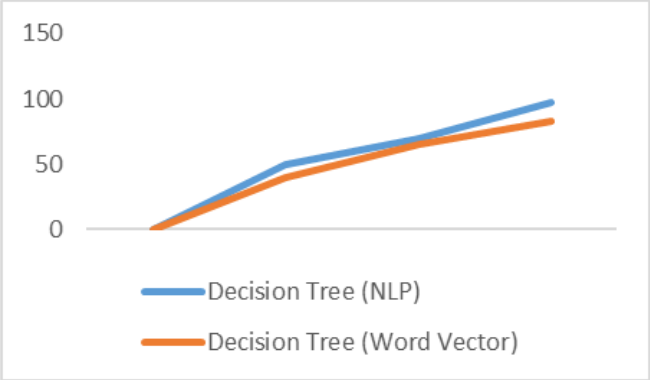

Fig. 11: URL based Phishing Evaluation

According to above graph, the results for the machine learning algorithm used for detecting phishing websites for the decision tree is 97.01% accurate with NLP feature and 82.47% accurate with World Vector.

C. Phishing – E-mail Attack

As per the result of ROC (Receiver Operative Characteristics – a graphical comparison plot between sensitivity and specificity) curve which is used to plot the result of TP (True Positive) vs FP (False Positive) which identify all the positive examples and is a perfect classifier to classify positive cases and negative cases efficiently. The accuracy of the test depends on the classification of the group and is measured by the area under the ROC curve. Through which the SVM achieved the best result with an accuracy of 98% whereas the Biased Support Vector machine and the CNN algorithm found to be same with 97% accuracy. The accuracy of K-means and phishing dataset is a clustering technique which found to be 90.68%.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

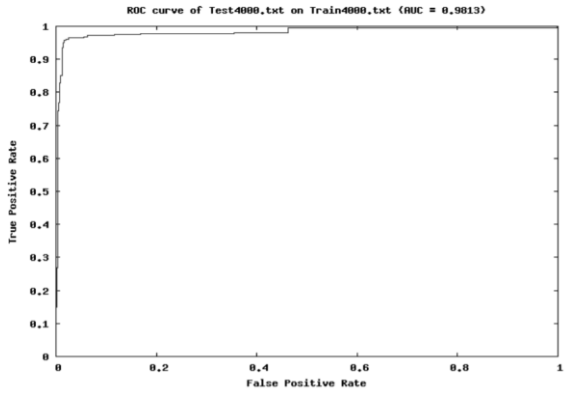

Fig. 12: Mail based Phishing Evaluation

VIII. CONCLUSION

Thus, the result obtained in the case study processing some Phishing Attacks provided through PhishTank and the Tamper Script obtained from SQLMap were successfully prevented by the Web Application Firewall (ModSecurity) and access to the website or the database server was forbidden. The Machine Learning approach is an efficient way to train the WAF to prevent the new kinds of attacks and protect the security and information of user. The result suggests the performance of the algorithm used in preventing Phishing Attack, Decision Tree and other approaches described in Section 6 is efficient and provides a good mechanism to identifying attack patterns. In the Future work we will investigate automated methods to produce effective patch for the Web Application Firewall and automate its repairing process.

REFRENCES

1. Abdul Razzaq, Ali Hur, Sidra Shahbaz, Muddassar Masood, H Farooq Ahmad- "Critical Analysis on Web Application Firewall Solutions", Issue, 2013
2. Abdulrahman Alzahrani, Ali Alqazzaz, Huirong Fu, Nabil Almashfi, Ye Zhu- "Web Application Security Tools Analysis", Issue, 2017.
3. Dennis Appelt, Cu D. Nguyen, Lionel Briand- "Behind an Application Firewall, Are We Safe from SQL Injection Attacks?", Issue, 2015
4. Dennis Appelt, Annibale Panichella, Lionel Briand- "Automatically Repairing Web Application Firewalls Based on Successful SQL Injection Attacks", Issue, 2017
5. Sandeep Kumar1, Renuka Mahajan2, Naresh Kumar3, Sunil Kumar Khatri- "A Study on Web Application Security and Detecting Security Vulnerabilities", Issue, 2017
6. Dennis Appelt, Cu D. Nguyen, Annibale Panichella, and Lionel C. Briand, *Fellow, IEEE*- "A Machine-Learning-Driven Evolutionary Approach for Testing Web Application Firewalls", Issue, 2018
7. Ram Basnet, Srinivas Mukkamala, Andrew H. Sung- "Detection of Phishing Attacks: A Machine Learning Approach" Issue, 2008
8. Yasin Sönmez, Türker Tuncer, Hüseyin Gökal, Engin Avci- "Phishing Web Sites Features Classification Based on Extreme Learning Machine" Issue, 2018
9. Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, Banu Diri -" Machine learning based phishing detection from URLs", Issue, 2018
10. A. D. Brucker, L. Brugger, P. Kearney, and B. Wolff- "Verified firewall ¨ policy transformations for test case generation", Issue 2010.
11. D. Appelt, N. Alshahwan, and L. Briand- "Assessing the impact of firewalls and database proxies on SQL injection testing", Issue 2013.
12. D. Gupta, J. Bau, E. Bursztein, and J. Mitchell- "State of the art: Automated black-box web application vulnerability testing", Issue 2010.
13. W. Halfond, J. Viegas, and A. Orso- "A classification of sql-injection attacks and countermeasures", Issue 2006.
14. Y.F. Li, P. K. Das, and D. L. Dowe- "Two decades of web application testing: A survey of recent advances", Issue 2014.
15. Tajpour, Maslin Masrom, Mohammad Zaman Heydari, Atefeh, and Suhaimi Ibrahim- "SQL injection detection and prevention tools assessment", Issue 2010.
16. M. Curphey and R. Arawo- "Web application security assessment tools", Issue 2006.
17. W. Stallings- "Network security essentials: applications and standards. Pearson Education India", Issue 2007.
18. E. Al-Shaer, R. Boutaba, H. Hamed, and M. Hasan- "Conflict classification and analysis of distributed firewall policies", Issue 2005.
19. Jim Beechey- "Web Application Firewalls: Defense in Depth for Your Web Infrastructure", Issue 2009.

AUTHORS PROFILE

Dr.J.Jagadessan, Head of Department, Department of CSE, SRM Institute of Science and Technology, Ramapuram, Chennai

Akshat Shrivastava, Student, Department of CSE, SRM Institute of Science and Technology, Ramapuram, Chennai

Md Arman Ansari, Student, Department of CSE, SRM Institute of Science and Technology Ramapuram, Chennai

Laxmi Kanta Kar, Student, Department of CSE, SRM Institute of Science and Technology, Ramapuram, Chennai

Mukul Kumar, Student, Department of CSE, SRM Institute of Science and Technology, Ramapuram, Chennai

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)