An Enhanced Malicious Web Content Detector and Preventor System
Bunmi Deborah Millennial-Oriagbo
M.Sc (UPH), PGD (UPH), B.Sc (UI), ND (Fed. Poly Ede).
Department of Computer Science
Federal Polytechnic of Oil and Gas Bonny Rivers State, Port Harcourt Rivers State, Nigeria

Abstract:
Malicious web contents prevention and detection is now an important topic in the discipline of information security owing to the emerging powerful techniques often utilized by the hackers to compromise user computer system. Some or few web pages are benign midst the majority of web pages that contains malicious web content. Most anti-virus packages in use are based on the use of signature-based access and these are not able to reveal camouflaged malicious HTML codes. Hence, this project work proposed a malicious web page detector and preventive measure using CORTICAL ALGORITHM approach of machine learning. This project examines the behaviour of malicious webpages, compares the existing Naïve Bayes algorithm used in the detection of malicious web content with the proposed Cortical algorithm used for the design and implementation of this new system. Experimental results reveals that the new system is not just capable of detecting the malicious web contents on webpages perfectly, it’s also capable of removing iFrames and blocking Popups meant to cause distractions and to frustrate the users efforts while working on the computer system. This research is the first of its kind to effectively block Popups and remove iFrames with the use of CORTICAL ALGORITHM.

I. INTRODUCTION

Background to the Study
A malicious site page is a webpage that contains harmful substance that can destroy a customer side PC framework. These malicious site may attracts a weapon by cybercriminal to abuse different security dangers, for instance, phishing, drive-by download, data taking, and spamming. As indicated by Kaspersky L. (2017), malware is a type of PC program meant to taint a real client's PC and cause hurt on it in various ways. As the decent variety of malware is expanding, hostile to infection scanners can't fulfill the requirements of security, Leading to a large number of hosts being assaulted. Kaspersky L. (2017), at the same time expressed that around 6 563 145 distinct hosts were assaulted, and 4 000 novel malware items were identified in 2015. Juniper H. (2016) predicts the expense of information ruptures to increase to $2.1 trillion universally by 2019.

How malware functions
Malware suppliers utilize an assortment of visible and unclear intentions to populate malware that attack gadgets and systems. For example, these malicious contents can be conveyed to a computer system via USB drive or can populate over the web through drive-by downloads, that consequently download malicious programs to computers without the client's endorsement or information. Another normal kind of malware is Phishing attacks which camouflages messages as genuine messages and contains malignant connections or connections that can carry executable malware to the unsuspecting clients. Powerful malware assaults frequently encompass the use of an order and-control server that enables danger on-screen characters to communicate with the tainted frameworks, expel delicate information and even remotely control the bargained gadget or server. Ongoing strains of malware incorporate new methods and inovative procedures that are aim to misdirect the clients and also the security heads and hostile to malware items too. A portion of these avoidance methods relies on straightforward strategies, for example, utilizing web intermediaries to shroud malicious traffic or source IP addresses. All the more, dominant dangers incorporate polymorphic malware, which continuously change its basic code to maintain a strategic distance from discovery from mark based recognition devices, hostile to sandbox procedures, which permit the malware to recognize when it is being examined and defer execution until after it leaves the sandbox, and document less malware, which dwells just in the system RAM so as to avoid being found. Distinctive malware types contain interesting qualities and attributes.

Types of malware include:  
Virus: This is the most widely recognized kind of malware that can execute itself and distribute by contaminating different programs or records.
Worm: Worm can self-recreate with the absence of a host program and commonly spreads with no human cooperation or orders from the malware writers.
Trojan horse: This is intended to show up as an authentic program so as to access a system. When initiated after installation, it can execute their vindictive capacities.
Spyware: Is made to gather data and information on the gadget client and watch their movement without their insight.
Ransomware: This is intended to attack a client's system and encrypt the information. Cybercriminals then request a payment installment from the unfortunate casualty in return for decrypting the system information.
Rootkit: Rootkit is created to get chairman level access to the unfortunate casualty's framework. Once introduced, it gives danger on-screen characters root or favored access to the framework.
Backdoor virus or remote access Trojan (RAT): Secretly makes a secondary passage into a contaminated system that enables threat entertainers to remotely get to it without alarming the client or the system security programs.
Adware: This is utilized to follow a client's program and download history with the goal to show pop-up or banner
advertisements that tempt the user into making a buy. For instance, a advertisers may utilize cookies to follow the website pages a client visits to all the more likely objective publicizing.

Keyloggers: Is additionally called system screens, are utilized to see almost everything a client does on their PC. This incorporates messages, opened site pages, programs and keystrokes.

Mobile malware
Cell phones are likewise inclined to malware assaults that may give access to the gadget's parts, for instance, the camera, mouthpiece, GPS or accelerometer. Malware can infect a cell phone if the client downloads an informal application or in the event that they click on a malicious link from an email or instant message. A cell phone can likewise be reached via a Bluetooth or Wi-Fi connection. Malware is discovered exorbitantly more on gadgets that run the Android OS nearly to iOS gadgets. Malware found on Android gadgets is normally downloaded via applications. One great signs that an Android gadget is contaminated with malware include irregular increments in information use, a rapidly scattering battery charge or calls, messages and messages being sent to the gadget contacts without the client's learning. Additionally, when a client gets a message from a perceived contact that appears to be suspicious, it might be from a sort of a versatile malware that distributes between gadgets. Gadgets with apple iOS are infrequently attached with malware in light of the fact that Apple cautiously vets the applications sold in the App Store. Then again, it is as yet feasible for an iOS gadget to be contaminated by exposing an obscure connection found in an email or instant message. iOS gadgets then result to be progressively helpless when jailbroken.

II. LITERATURE REVIEW

Review of Models for Malicious Detection and Prevention on Web Site
Malicious detection and prevention of site is one of the hot research themes in the field of security. The significance is assessed from these insights which demonstrate that Google finds 9,500 new malignant Web locales daily. This is likewise significant on the grounds that it keeps the client from being victims of attacks, for instance, Phishing attacks, Drive-by-download attacks, spam attack, Click-jacking, Plug-In and Script-Enabled Attacks, Mal-advertising. Malicious web identification is seen as a way of recognizing those web URLs and website pages which can cause trade off in client security and influence the clients (Karl, 2002).

Without a doubt, malicious detection on the Internet (cybercrime) have expanded at a lot quicker rate and with high multifaceted nature. Cybercrime assaults includes online frauds, breaking into the system, phishing assaults, DNS harming, malware assaults, information stealing, spamming, tricks, extorting. An ongoing case of cybercrime is Sony Cyber Hack through which the PC at the corporate network of Sony Pictures were ruptured and taken disconnected by a malware-based attacks.

Fraites, T. (2008) archived that intrusion detection and prevention system are programming or equipment parts that screen PC frameworks and break down occasions happening in

them for indications of interruptions. Because of generally utilization of and multifaceted nature of PC frameworks, it is hard to give a totally secure PC framework. Then again, there are various security system and intrusion detection and prevention system that address various parts of PC security.

Fraites, T. (2008) archived that an intrusion detection and prevention system or framework activities for malicious activities on website pages or approach infringement and produces reports to an administration station. Intrusive events to PC systems are expanding because of the loving and reception of internet and local area network computerized hacking instruments and technique. PCs are gaining the chance to be increasingly more exposed to attacks because of its wide spread system network.

History
Malware was first used by PC researcher and security analyst Yisrael Radai in 1990. Meanwhile, malware existed few time before this time. Creeper Virus was the most popular Virus in 1971, that was made as an investigation by BBN Technologies engineer Robert Thomas. Creeper was intended to attack centralized servers on ARPANET. The program did not adjust capacities, or take or erase information, the program was then moved beginning with one centralized computer then onto the next without consent while showing a print message that read, "I'm the creeper: Catch me in the event that you can." Creeper was later modified by PC researcher Ray Tomlinson, who included the capacity to self-reproduce to the infection and made the principal realized PC worm. The idea of malware began with the innovation business, different instances of infections and worms started to show up on Apple and IBM PCs in the mid-1980s before winding up broadly known after the presentation of the World Wide Web and the business web during the 1990s.

Roles of Malicious Detection System
According to Fuchsberger, (2005), the role of the detector and preventer is to find and dispense with unneeded data from the review trail. It then displays either a manufactured views on the security-related moves created during normal utilization of the framework, or the engineered views on the present security state of the framework. A choice is then made to assess the probability that these activities or this conduct can be considered as side effects of an intrusion or vulnerabilities. A preventive measure part would then be able to make restorative move to either keep the activities from being executed or change the condition of the system back to a protected state (Fuchsberger, 2005). An interruption recognition framework assembles data about a data framework to play out a conclusion on the security status of the last mentioned. The point is to find ruptures of security, endeavoured breaks, or uncovered vulnerabilities that could prompt potential breaks (Fuchsberger, 2005).

Modules, for example, ActiveX is notable for their powerlessness to perversity dynamic substance. Moreover, malicious dynamic substance can take a secret phrase or stick and later get to a site with classified data while causing it to seem like it was gotten to by the approved client. This can make it even hard to follow whether malicious dynamic substance is used up in the attack.

Framework of Cortical on multi-class Malicious URLs Detection System: Cortical Algorithm (CAs) rose as a naturally
roused methodology, displayed after the human visual cortex, which stores arrangements of examples in an invariant structure and reviews those examples autoassociatively.

**Cortical Algorithm Structure**

Human brain is comprised of a six-layered structure comprising of an exceptionally enormous number of neurons emphatically associated by means of feed forward and feedback connections. A significant property of the neocortex is its basic and useful consistency: all units in the system appear to be comparable, and they play out a similar essential activity. Precisely like brain architecture, CA design has minicolumns of shifting thickness (Edelman and Mountcastle 1978). A minicolumn is a gathering of neurons that offer the equivalent open field: neurons having a place with a minicolumn are related with the equivalent sensory input region. The minicolumn is the essential structure in a cortical system network, in contrast with neurons in an old style ANN. A relationship of minicolumns is known as a hypercolumn or layer (in what pursues, the terms segment and minicolumn are utilized reciprocally). Associations in a CA system happen in two ways: on a horizontally, between segments in a similar layer, and vertically, between sections of sequential layers. In spite of the fact that associations between nonconsecutive layers are available in the human cortex, these associations are evacuated in CA, for effortlessness purpose.

**Iframe**

Iframe is a html a html label utilized in including outer site into your website(webpage).

**Iframe infection**

An Iframe virus is a malicious code that infect pages on sites. These are considered as type of malware. A large portion of them use iframe html code, causing harm by infusing iframe labels into the site. Code might be infused into html, php, records. The infection may make its essence known by checking for landing page records, for example, index.php, index.html or default.html and infuse the iframe code in them. The iframe code is typically found close to the start on the website page. They may likewise contaminate through subjects or formats of substance the executives frameworks. The virus will likewise alter .htaccess and hosts documents, and make images.php records in registries named 'pictures'. The contamination isn't a server-wide endeavor, it will just infect locales on the server that it has passwords to. This ongoing flood in bargained web servers has produced exchanges in online discussions and web journals. Web malware diseases hurt organizations; google, firefox, web adventurer and against infection organizations boycott contaminated destinations, organizations lose income and locales endure harm to their image and notoriety. An iframe infection is a sort of badware. "badware makers are always growing new, innovative approaches to introduce badware onto your computer". badware dispersion has been extended past conventional channels like email infections to harder-to-stay away from techniques like mechanized "drive-by downloads" that are propelled by traded off site pages.

**Iframe variations**

Here and there iframe variations come as JavaScript. iframe labes may not be found in plain content in the source since it is encoded. On the off chance that the encoded content code is decoded, it will contain code to conjure iframe by means of JavaScript.

**Iframe Injection Website Attack and Tips to Clean the Infection:**

Most popular online attacks that happens all the time has to do with a huge ammount of real sites being hacked with the Iframe code infusion assault bringing about Cross Site Scripting (XSS) or quiet redirections to malevolent sites. This implies unsuspicious guests get contaminated with a malware infection when they visit those genuine sites. The result is that an "Iframe html code" is infused toward the end (for the most part) of index.php or index.html documents of the real sites. This iframe code inserts into the genuine site a malignant code which introduces an infection to the guest’s PC or attempts to take delicate data. The <iframe> html tag is utilized to install content from other site into the present page. Typically an infused iframe code resembles the accompanying:

```
<iframe src="http://some-vindictive web-url" width=1 height=1 style="display:none"></iframe>
```

**Challenges in the Malicious Web Content Detection and Prevention Techniques**

Majority of the present methods used in detecting and preventing malicious site depend on their center procedures for a notable attack. Hence, the attackers continue creating amendments in the existing methodology and acquaints new systems with be implanted in site page. The current techniques depends upon the fixed arrangement of highlights yet the aggressor continues making amendments in the current features and furthermore presents new features. Thus, the discovery and anticipation strategies are not equipped for distinguishing the new attacks. This requires that the detection methods be improved. The various strategies like signature based, feature based and behavior based methodologies utilized in detecting malicious site and contents are confronting these confinements because of modern attacks. In accordance with the constraints, the various existing features are not adequate for the detection of malicious sites. For instance existing methodologies are not fit for distinguishing malicious sites based on the area name because the attackers often changes the space. Besides, none of the component gathering strategies can gather the rising highlights. The current detection techniques suffers false negativities. Thus, there should be a better way to solve this problems.

**Iframe**

Iframe is a html a html label utilized in including outer site into your website(webpage).

**Iframe infection**

An Iframe virus is a malicious code that infect pages on sites. These are considered as type of malware. A large portion of them use iframe html code, causing harm by infusing iframe labels into the site. Code might be infused into html, php, records. The infection may make its essence known by checking for landing page records, for example, index.php, index.html or default.html and infuse the iframe code in them. The iframe code is typically found close to the start on the website page. They may likewise contaminate through subjects or formats of substance the executives frameworks. The virus will likewise alter .htaccess and hosts documents, and make images.php records in registries named 'pictures'. The contamination isn't a server-wide endeavor, it will just infect locales on the server that it has passwords to. This ongoing flood in bargained web servers has produced exchanges in online discussions and web journals. Web malware diseases hurt organizations; google, firefox, web adventurer and against infection organizations boycott contaminated destinations, organizations lose income and locales endure harm to their image and notoriety. An iframe infection is a sort of badware. "badware makers are always growing new, innovative approaches to introduce badware onto your computer". badware dispersion has been extended past conventional channels like email infections to harder-to-stay away from techniques like mechanized "drive-by downloads" that are propelled by traded off site pages.

**Iframe variations**

Here and there iframe variations come as JavaScript. iframe labes may not be found in plain content in the source since it is encoded. On the off chance that the encoded content code is decoded, it will contain code to conjure iframe by means of JavaScript.

**Iframe Injection Website Attack and Tips to Clean the Infection:**

Most popular online attacks that happens all the time has to do with a huge ammount of real sites being hacked with the Iframe code infusion assault bringing about Cross Site Scripting (XSS) or quiet redirections to malevolent sites. This implies unsuspicious guests get contaminated with a malware infection when they visit those genuine sites. The result is that an "Iframe html code" is infused toward the end (for the most part) of index.php or index.html documents of the real sites. This iframe code inserts into the genuine site a malignant code which introduces an infection to the guest’s PC or attempts to take delicate data. The <iframe> html tag is utilized to install content from other site into the present page. Typically an infused iframe code resembles the accompanying:

```
<iframe src="http://some-vindictive web-url" width=1 height=1 style="display:none"></iframe>
```
FUNCTION OF MALWARE (Malicious Contents) on Webpages

i. Recover all put away username and secret word on internet browser.
ii. Hinders site execution and the internet browser all in all.
iii. Overabundance popup on the page

III. MATERIALS AND METHODS

Existing System Analysis

The current system was a Phishing URL detection system utilizing Naïve Bayes. These current methodology contrasts the suspicious website with the genuine one site by utilizing different highlights and on the ground that the difference is more than the predefined limit esteem, it is proclaimed phishing if the threshold is exceeded. Existing System catches images (screen capture) from suspicious URL. It at that point contrasts the given images and the put away images in the database utilizing ImgSeek (C. Jacobs, 1995), in the event that it will finds the comparative pictures in the database. After picture correlation, the framework thinks about the area name. On the off chance that the space name is absent in the database, it implies that the info URL is extraordinary. By then, the framework pronounces the given URL phishing. In the event that there is no picture in the doubts site whose likeness is more prominent than the edge esteem, at that point the proposed procedures return the outcome as obscure and register the picture in the database.

![Figure 3.1: Architecture of the Existing System - Bayesian model for phishing detection (Ankit Kumar Jain, January 2017).](image)

Algorithm of the Existing System

1. imgSeek(ImgSuspicious, ImgLegitimate, ImgPhish, ImgUnknown)
2. if (Similarity(ImgSuspicious, ImgLegitimate) > threshold)
3. if (domain(ImgSuspicious) = domain(ImgLegitimate)) then return("Legitimate")
4. else return("Phishing")
5. end if
6. else
7. store Database({ImgSuspicious, domain(ImgSuspicious), Unknown})
8. end if

Analysis of the Proposed System

The proposed system is Malicious site pages detection and prevention system utilizing cortical algorithm. The propose system utilizes the cortical algorithm to prepare some dataset and limit was set in term of malicious content and ordinary substance. The dataset is spoken to as a vectors and the estimations of the vectors as being set has 1.0 as would be expected characters while 5.0 is set as the vindictive threshold. The program show level of malicious attacks and how cut off it has been influencing the PC system in a quantities of popups. Every malware demonstrates diverse popups relying upon the record or report it has attack.

Algorithm of the Proposed System

1. Initialization input training samples; then digitizing and normalizing the input data;
2. Reducing the dimension of the data on the browser to indicate two set of vectors
3. Input vector with dimensionality reduction, network parameter to initialize the classifier;
4. Set the layer i=5;
5. Train the network layer by layer according to cortical learning rules, then save the result including the weights and biases;
6. If i<=max layer, set i=i+1; when i > max layer, do the supervised learning
7. Input the dataset test samples into the trained classifier to detect malicious code and the normal code.

Advantages of the Proposed System

Coming up next are a portion of the advantages of the proposed system:

- Easy to use: The new proposed system is easy to use by internet user when compared to the existing system.
- It detect and prevent the malicious contents in the new and existing websites.
- It partition the output into different rules which aid in identifying the normal and the malicious data.
User friendly and does not need an expert to operate. The point of the structure stage is to design an answer of the issue determined by the prerequisite record. This stage is the initial phase in moving from issue space to the arrangement area. Framework configuration is probably the most basic factor influencing the nature of the product, and majorly affects the later stages, especially testing and support. Each undertaking requires a structure so as to give the client mandates. In this venture, we utilized UML and calculation to make this program simple. Framework configuration is the way toward characterizing the engineering, segments, modules, interfaces, and information or a framework to fulfill indicated necessities.

IV. DISCUSSION OF RESULT/COMPARISON OF THE EXISTING AND PROPOSED SYSTEM RESULT.

The current methodologies contrast the suspicious site and the relating benign site by utilizing different features and in the ground that the similarity is more than the predefined threshold, at that point it is considered phishing. Existing System catches the images (screenshots) from suspicious URL. It at that point contrasts the given images and then put away images in the database utilizing ImgSeek, on the off chance that it will finds the comparable pictures in the database. After images correlation, the system compares the domain name. If the domain name is not present in the database, it means that the input URL is different. At that point, the system declares the given URL phishing. If there is no image in the suspicions website whose similarity is greater than the threshold value, then the proposed techniques return the result as unknown and register the image in the database. The proposed system is also a single algorithm for detection and prevention approach. The system algorithm detection threshold is set to be 5%. The algorithm scan both the existing URL domain names and any new URL domain names for detection and prevention purposes. The new system has high detection rate with other features capabilities not inherent in the existing system.

Software Methodology

Waterfall model

The Waterfall Model is a breakdown of venture activities into direct successive stages, where each stage relies upon the expectations of the past one and relates to a specialization of errands. The methodology is normal for specific territories of building plan. In programming development, it will in general be among the less iterative and adaptable methodologies, as forward movement streams in to a great extent one heading ("downwards" like a cascade) through the period of build, presentation, analysis, plan, development, testing, organization and support. The cascade development model began in the assembling and development industry, where the exceptionally organized physical situations imply that structure changes turned out to be restrictively costly much sooner in the improvement procedure. At the point when originally received for programming advancement, there were no perceived option for learning based innovative work.

WHY CHOOSE WATER FALL

Waterfall depends on groups following a succession of steps and failing to move forward until the past stage has been finished. This structure is fit for litter activities with expectations that are anything but difficult to characterize from the start. Ben Aston from The Digital Project Manager clarifies, "Cascade is for the most part viewed with some hate as a wasteful and out of date customary task the board approach. Be that as it may, Waterfall can be a helpful and unsurprising methodology if necessities are fixed, all around recorded, and clear, if the innovation is comprehended and develop, if the venture is short, and if there's no extra worth picked up from 'going deft.'

A Waterfall approach can really give progressively unsurprising final product to spending plan, course of events, and scope."Here's an inside and out take a gander at what the Waterfall system does best.
Choice of Programming Language
This simply refers to the type of programming language used in the proposed system. In this proposed system the programming language used is JavaScript. Reason been that javascript can be used as a client side scripting language. The javascript code can be written into HTML pages that are often used by most web browser applications. New system is developed on the platform of web page base used.

V. RESULTS AND DISCUSSION

Hardware Requirements
Hardware requirements are listed below:
- Color Monitor
- 512MB Random Access Memory (RAM Size)
- 1 Gigabyte Hard Disk Drive
- Keyboard
- Mouse

Software Requirements
The software requirements for the installation of the system are listed below:
- Web Browser Software Application
- JavaScript
- Mysql Database Server
- Xampp

Documentation and Implementation
Having confirmed that the proposed system meets the objectives of the project, the implementation phase begins. Implementation is the stage of a project during which theory is turned into practice. The proposed malicious web pages detection and prevention is implemented by installing JavaScript integrated development environment. Also Xampp server was installed for the smooth running of the web page on the web site.

Running the Application.
Lunch the JavaScript IDE on the computer, and it will open the malicious detection program. Right click on the program and click on run to execute the program. Then lunch the web site on the computer system to link the web site to the program and also copy the link of the site into the program to verify the content of the malicious code that have attack the data and files in the web page. Then click on browser to refer to each document and file in the web site. The application will detect all the component of the file and document in the site and give result about the web page.

Main Menu
The main menu is a number of options made available to the user of a computer application to help the user find information or execute a program function. The main menu for the proposed system showing the various layouts of the various user interfaces is illustrated below:

VI. SUMMARY, CONCLUSION AND RECOMMENDATIONS

Summary:
Malicious Web pages are increasingly spread while accessing the website on our different computers daily. However, these malicious content or code that attack web pages on the web browser goes directly into the computer local disk and stored in different files or document and also create a storage space, thereby affecting and damaging different data in the computer. These malicious contents that attack the web pages also cause different types of computer virus that infect the computer system to mal function. However, in spite of significant advances in processor power and bandwidth, the browsing experience on different computer devices is considerably different when the computer is been affected by these threat. The advancement in computers have made computer networks and other smart devices to increase the amount of services that are available on
the Internet with many people accessing the website through web applications. A number of these web applications provide convenient services to users, which include online commerce, communicating through social network application and services or surfing for information online.

**Conclusion**

A malicious web page or codes are malware that contained harmful content that can destroy a client-side computer system. This type of attack is termed web-based client-side attack. The attack is delivered as part of the web page itself and is designed to exploit client-side vulnerabilities such as flaws in the implementation of browser functionality, interpreters of active content within WebPages or scriptable client-side components such as HTML components. This new system is capable of preventing the pop ups meant to course unnecessary distractions when users operate the computer system and to also prevents iFrames viruses that secretly penetrate and infect the users system.

**Recommendations**

We recommend that this software should be packaged as an Addons and be made available on various web browser applications stores so that users can add it to their web browsers for detection and prevention of malicious contents purposes.

**Contribution to Knowledge**

We have been able to remove iFrame and pop ups from webpages automatically with JavaScript whether offline or online. This is the basis of webpages detection and prevention.

**VII. REFERENCES**


[7]. Karl A. McCabe … US7607171B1 (en) *, 2002-01-17, 2009-10-20, Avinti, Inc … US20090271867A1

[8].(en)*,2005-12-30, 2009-10-29, Peng Zhang, Virtual machine to detect malicious code …


[20]. [https://moz.com/top500](https://moz.com/top500)