MALWARE CLASSIFICATION USING MACHINE LEARNING ALGORITHMS AND TOOLS

Mr. Chavan Vikram
Computer Technology, Late G. N. Sapkal College of Engineering, Nashik, India.

Mr. Daund Hritik
Computer Technology, Late G. N. Sapkal College of Engineering, Nashik, India.

Mrs. Dhanagar Gayatri
Computer Technology, Late G. N. Sapkal College of Engineering, Nashik, India.

Mrs. Patil Harshada
Computer Technology, Late G. N. Sapkal College of Engineering, Nashik, India.

Abstract

The explosive growth of malware variants poses a major threat to information security. Malware is the one which frequently growing day by day and becomes major threats to the Internet Security. According to numerous increasing of worm malware in the networks nowadays, it became a serious danger that threatens our computers. Networks attackers did these attacks by designing the worms. A designed system model is needed to defy these threats, prevent it from multiplying and spreading through the network, and harm our computers. In this paper, we designed a classification on system model for this issue. The designed system detects the worm malware that depends on the information of the dataset that is taken from website, the system will receive the input package and then analyze it, the Naïve Bayesian classification technique will start to work and begin to classify the package, by using the data mining Naïve Bayesian classification technique, the system worked fast and gained great results in detecting the worm. By applying the Naïve Bayesian classification technique using its probability mathematical equations for both threat data and benign data, the technique will detect the malware and classify data whether it was threat or benign.

Keywords - *Malware, Machine Learning, Classification, Naïve Bayesian*

I. INTRODUCTION

With the rapid development of the Internet, malware became one of the major cyber threats nowadays. Any software performing malicious actions, including information stealing, espionage, etc. can be referred to as malware. Kaspersky Labs define malware as “a type of computer program designed to infect a legitimate user's computer and inflict harm on it in multiple ways.In this world of digitization ,one of the most immediate threats to one's professional data and personal data is malicious malware executable. Intruders & hackers are using various new approaches of intruding different type of malware in existing software's like polymorphic metamorphic are very difficult to recognize or categorized accurately. Lot of efforts are required to analyze huge number of malware samples manually. Malware word defines from Malicious Software. Malware is a malicious code that affects the user system or computer and intently harms the computer by an attacker. Malware is variant forms which are a virus, Trojan, backdoor, root kits, ransom ware, worm, botnet, spyware, adware, key loggers, etc., and there is a wide range of their families are existing and massively growing on the internet daily.Malware have impacted a large number of computing devices. The term malware come from malicious software which are designed to meet the harmful intent of a malicious attacker. Malware can compromise computers/smart devices, steal confidential information, penetrate networks, and cripple critical infrastructures, etc. These programs include viruses, worms, trojans, spyware, bots, rootkits, ransomware, etc.Network security is an important branch of computer science that protect the stored data in the computers, which it connected together by one network. Recently, the knowledge of the network became developed and common in our world, network attackers are increasing every day, and their threats are evolving as well.Network security is a very important matter for foundations like universities, special projects and corporations. These foundations can supply many important functions for the countries safety.Nowadays, the online services are very popular for the users. The users now can communicate with each other and share information, and knowledge among each other. Now these services are less expensive and more cooperative by using the Information Technology (IT) associations, and Internet Service Providers (ISPs).Malware may put network at danger. Malware is a program can install in the network and electronic devices like computers, smart phones and tablets that connected in the network. It damages these devices by accessing it illegitimately and destroy its personal data and information; for an example: Adware could do the malicious work . Malware is the most dangerous threats to the networks. The malware could take many forms to do its attack, it always come as package and try to access to the network. Every day new types and forms of the malware are found. The malware programmers always make decisions about protecting their malware from anti malware programs like Kaspersky, McAfee, NOD, Norton and many anti viruses programs that we use in our PCs .The malware threats are too many; the most serious one is worm. The worm can replicate itself and spread through the network very fast. Nowadays, the world faces a major problem called worm especially the facilities and the network users. In spite of the detection techniques ability in detection it still have difficulties in detecting these worms . Here the detection techniques role comes. Detection techniques are the most effective defense against the malware in the network. The malware defenders are the anti-viruses these days, it can detect the malware signature and prevent it from doing its malicious work . In the Operating System (OS) keeping the safety, confidentiality and availability is very important and a hard task, because of the difficulties and impendence that the network faces in securing the data and information inside the network and keeping it safe from outside attacks, so it is very important these networks have a defense line against the outside attacks.Electronic devices such as computers and smart phones; the malware can affect them by a huge number of malware that spread in these devices.

II. LITERATURE SURVEY

In the past few years, researchers and anti-malware communities have reported using machine learning and deep learning based methods for designing malware analysis and detection system.

According to the survey [1] conducted by AV-Test Institute, it registers that everyday 350,000 new malicious code and potentially unwanted applications. Each malicious one is classified with respect to their behavior and saved accordingly by this institute and gives the malware statistics in 2018 is 847.34m malicious code is found and recorded and registered. Some of the malware attacks in history are Melissa was a macro embedded with a word file. When the user opens it the macro will execute and resend the virus to the first 50 people in the user's address book. It was designed by David L. Smith in 1999. Likewise, several malware attacks in history namely, My Doom worm in 2004, Stuxnet in 2010, wannacry in 2017. In this paper, we presented the literature work of previously existing works of malware detection classification using machine learning algorithms.

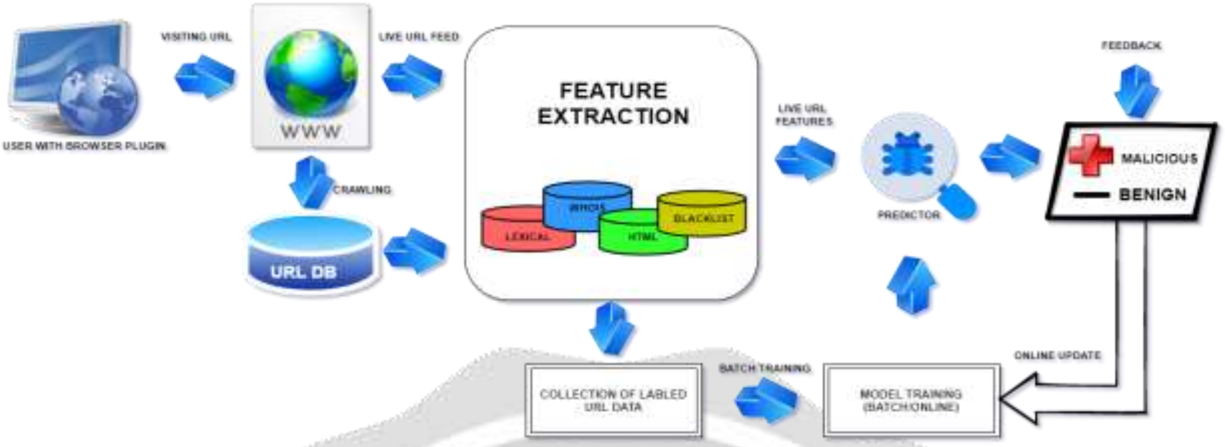
In 2013 Xiao, et al. proposed a system to detect new types of malware; they used API (Application Programing Interface) with OOA (Object-Oriented Associate) mining algorithm, which it used for association rules for detection atter. After many tests the proposed system showed great results and new different types of malware have been discovered [14]

Younghee Park, et al. presented a method to detect the malware that depend on the known behavior of graph that show the behavior of malware instances. This method uses clustering in detection; it clusters the set of separated behavior graphs and make a single behavior graph. The results included numerous detection rates with almost 0% of false alarms

In “Zero-day Malware Detection based on Supervised Learning Algorithms of API call Signatures”, the API functions were used for feature representation again. The best result was achieved with Support Vector Machines algorithm with normalized polykernel. The precision of 97.6% was achieved, with a falsepositive rate of 0.025. (Alazab, et al. 2011).

III. SYSTEM ARCHITECTURE

The system architecture is an innovative platform focused on detecting potentially risky websites. Thus, it is able to classify websites into risky or non-risky. For this purpose, it extracts knowledge from external web information sources and makes predictions when no information is available. In order to make these predictions, DOCRIW builds similarity measures to train ML algorithms, and uses optimization methods to select the best model and the proper parameters. Notice that the proposed approach refers to direct access of a website from users (i.e., when the users are trying to be scummed). Regarding the general architecture of the system, it presents four main modules (see Fig. 1): the Domains Extraction and Validation module, the Host-based Variables Extraction module, the Classification module and the Information Updating module. Besides, the system also holds a Graphical Interface and two databases: the Knowledge Base and the Machine Learning Model. The Graphical Interface is in charge of the interaction with users. The Knowledge Base is an ElasticSearch database [48] that organizes the knowledge collected from the Web Information Sources. The Machine Learning Model includes a classifier previously trained. Next, the rest of the modules are described.

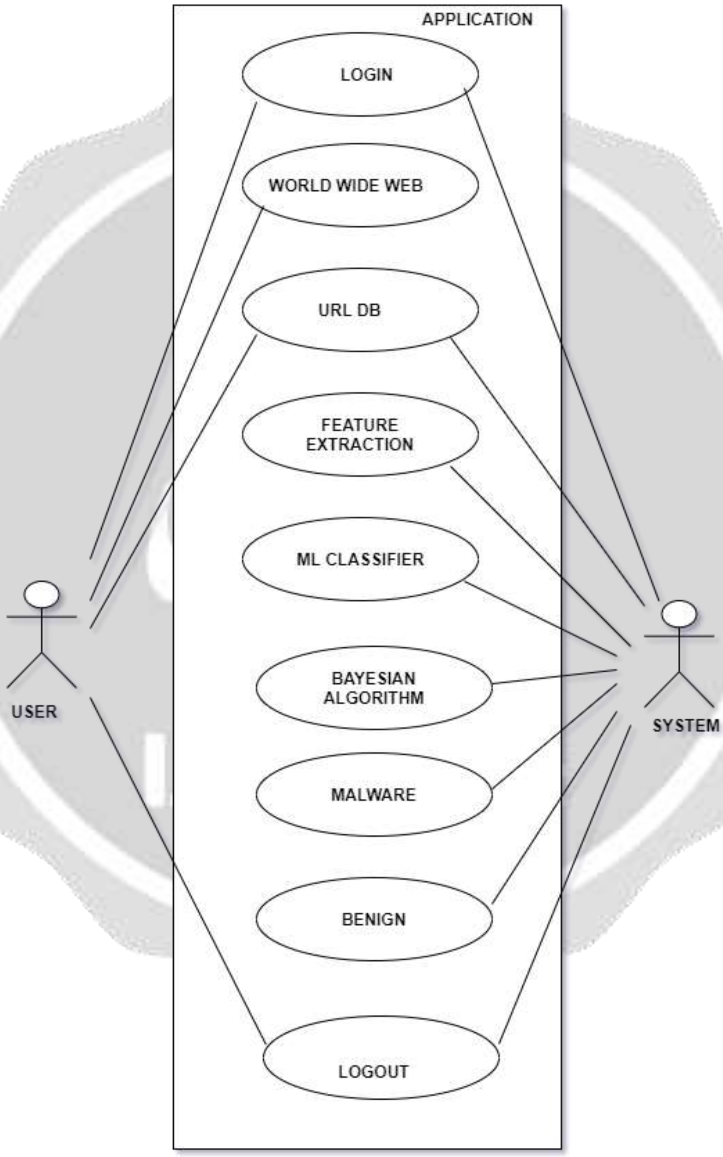


A. DOMAINS EXTRACTION AND VALIDATION MODULE This module processes URLs by extracting their corresponding domains and analyzing them. In order to achieve these tasks, it uses the Knowledge Base module to obtain the previously labeled risky domains. The module presents two components: the URL Analyzer and the Domains Evaluator (see Fig. 2). The first receives information from the Graphical Interface and acts in response to the requests made by users. The information provided by the Graphical Interface can be entire URLs or domains previously preprocessed. The URL Analyzer evaluates the proposed domain in both situations. Thus, it checks if the domain is correct (i.e., status code equals to 200) and it detects possible redirections to landing pages. In this case, all the landing pages are included to be analyzed, extracting the associated domains. The Domains Evaluator component matches the obtained domains and the domains stored in the database. When matches are found, the reported domain is labeled as risky. When none of the domains are matched, the module sends the original domain to the Host-based Variables Extraction module.

B. HOST-BASED VARIABLES EXTRACTION MODULE The Host-based Variables Extraction module collects new host-based variables through the Whois API REST [49], which is part of the Web Information Sources. It provides information about city, country, creation date, expiration date and e-mail. Thus, this module characterizes the analyzed domain. Regarding the architecture of the module (see Fig. 3), it consists of two components: the Host-based Variables Collector and the Data Cleaning. The first one manages the information provided by the Whois API, building a dataset as output. The second one addresses the cleanup task unifying the results. For instance, the country abbreviations areadapted according to the ISO code [50], and possible mismatches between values are normalized (e.g., a city name with accent marks and the same city name without them)

C. CLASSIFICATION MODULE This module classifies domains into risky or non-risky labels when they are not found in the Knowledge Base. It uses the variables generated by the Host-based Variables Extraction module to feed the Machine Learning Model in order to obtain a predicted value for domains. The Machine Learning Model has been selected based on empirical results. The complete study to select the elements related to this model will be explained later. The model includes a definition of the similarity between domains, a LR algorithm, a threshold for the probability provided by the algorithm, and a reference set of domains. Regarding the architecture of the module (see Fig. 4), it consists of two components: the Similarity Creator and the Classifier. The first one calculates similarities between the new domain and any of the domains in the reference set, for each variable (i.e., domain name, city, country, creation date, expiration date and e-mail). The similarity based on domain name is calculated using the Levenshtein distance [51]. The other five similarities (corresponding to the host-based variables) evaluate whether two domains have the same value for the corresponding variable or not. For instance, for the country variable, the similarity is 0 when the two domains are hosted in two different countries, and it is 1 when the two domains are hosted in the same country. Next, a global similarity between the new domain and any of the domains in the reference set, is calculated as a weighted average of the previous similarities. These weights are provided by the Machine Learning ModeThe second component of the Classification module is the Classifier. The LR algorithm provided by the Machine Learning Model is fed with the vector of global similarities previously calculated to obtained a prediction (between 0 and 1) of the riskiness. Given a predefined cut-off probability threshold that maximize the overall performance, the domain is labeled as risky or non-risky.

D. INFORMATION UPDATING MODULE This module collects data from the Web Information Sources to update the information stored in the Knowledge Base. Thus, it obtains new reported malicious domains from AA419 [52] and MalwareURL.com [53], two public websites thatidentify risky domains and makes this data available as a public service. This task is periodically executed by updating former register with the new gathered information. This module stores the information in the Risky Domainsindex of the Knowledge Base module.

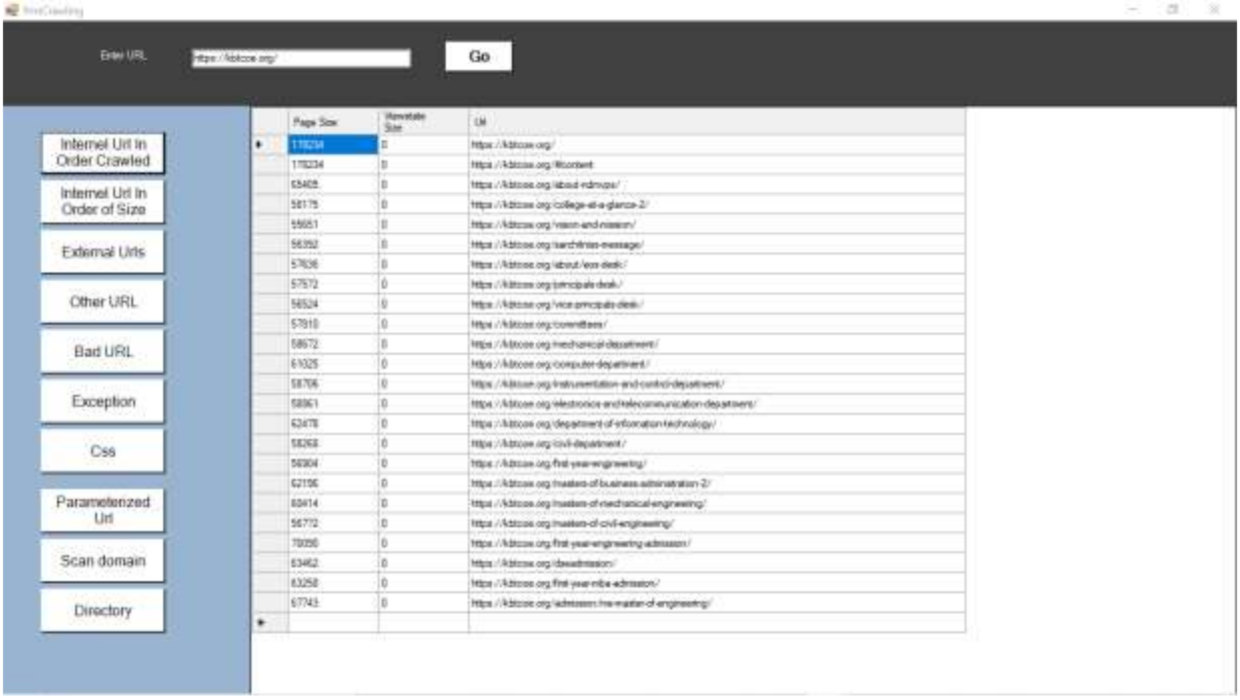


IV. TECHNOLOGY NECESSITY

First, we have done problem analysis and we recognize the users' problem of dealing with malicious websites, we have discussed and planned the solution to the problem by creating a system application. Different problems were faced while working on the solution, for which we conduct research on problems of our project, such as collecting information on different classifiers that classifies which type of virus is this. We have used the simple user interface easy for users to use. By upholding the ethics of application use, this application can be used by internet users. This software is created by cooperation and it is sustainable because it does not exist in physical form. To plan and build this project, we followed software engineering principles. We used the algorithm for concepts and languages for programming such as java. We followed the life cycle of a software development to create our project. We also mastered the important aspects of working as a team after undertaking all of these tasks and completing all the procedures.

V. RESULT

Apart from that application has several tabs such as Internal URL in order of crawled, Internal URL in order of size, External URL, Other URL, Bad URL, Exception, CSS, Parameterized URL, Scan Domain, Directory. Internal URL in order of Crawled tab is shown in fig 1.1 it shows the crawled URL's of a website.

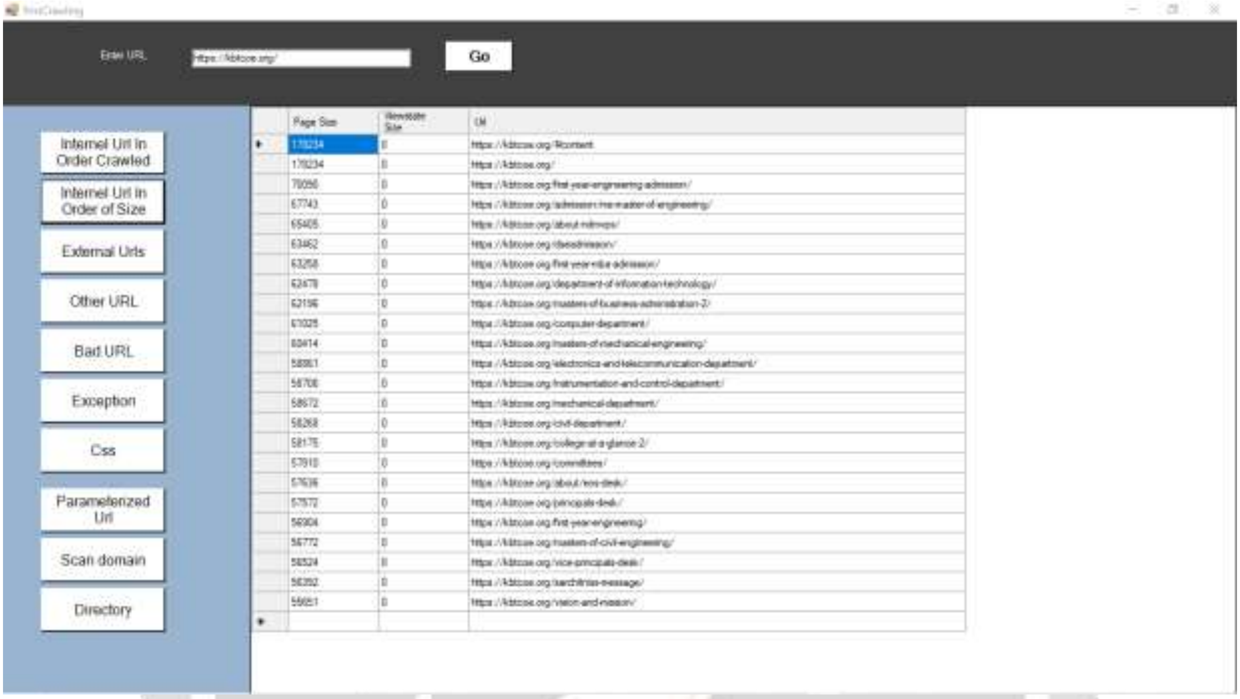


Enter URL https://kbtcoe.org/

Page Size	Viewstate Size	Url
170234	0	https://kbtcoe.org/
170234	0	https://kbtcoe.org/#content
65405	0	https://kbtcoe.org/about-ndmvps/
58175	0	https://kbtcoe.org/college-at-a-glance-2/
55651	0	https://kbtcoe.org/vision-and-mission/
56392	0	https://kbtcoe.org/sarchitnis-message/
57636	0	https://kbtcoe.org/about/ceo-desk/
57572	0	https://kbtcoe.org/principals-desk/
56524	0	https://kbtcoe.org/vice-principals-desk/
57810	0	https://kbtcoe.org/committees/
58672	0	https://kbtcoe.org/mechanical-department/
61025	0	https://kbtcoe.org/computer-department/
58706	0	https://kbtcoe.org/instrumentation-and-control-department/
58961	0	https://kbtcoe.org/electronics-and-telecommunication-department/
62478	0	https://kbtcoe.org/department-of-information-technology/
58268	0	https://kbtcoe.org/civil-department/
56904	0	https://kbtcoe.org/first-year-engineering/
62156	0	https://kbtcoe.org/masters-of-business-administration-2/
60414	0	https://kbtcoe.org/masters-of-mechanical-engineering/
56772	0	https://kbtcoe.org/masters-of-civil-engineering/
70050	0	https://kbtcoe.org/first-year-engineering-admission/
63462	0	https://kbtcoe.org/dseadmission/
63258	0	https://kbtcoe.org/first-year-mba-admission/
67743	0	https://kbtcoe.org/admission/me-master-of-engineering/

Fig. 1.1

Next tab is Internal URL in order of size it shows the page size of that website. This tab is shown in below fig 2.1

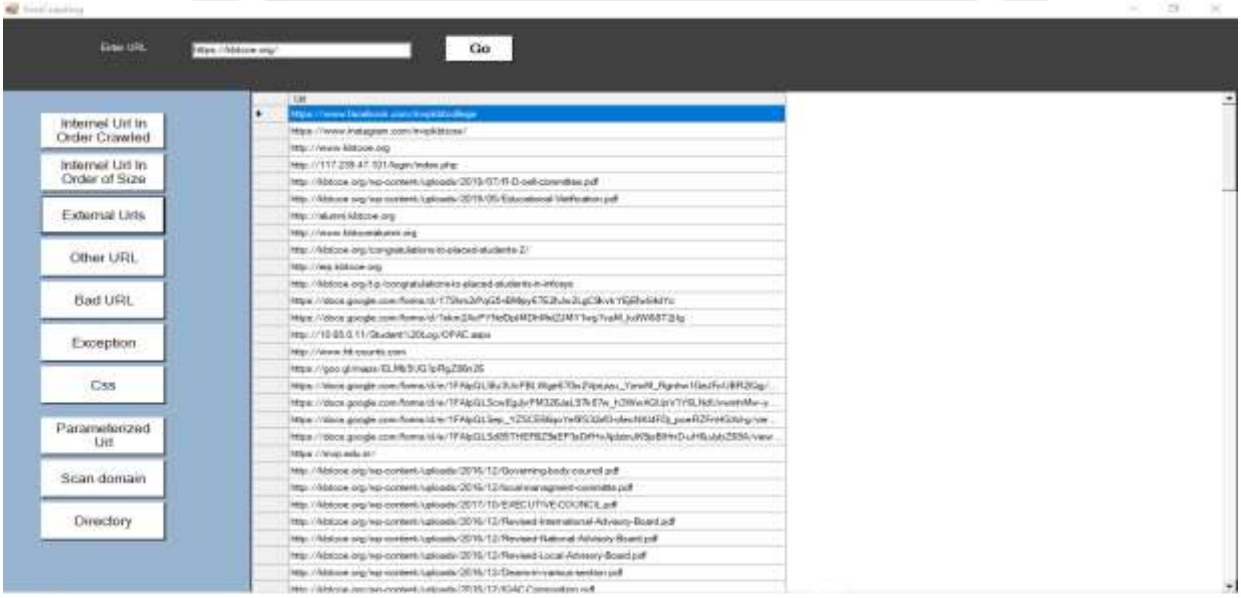


https://kbtcoe.org/

Page Size	Viewstate Size	Url
178234	0	https://kbtcoe.org/#content
178234	0	https://kbtcoe.org/
70056	0	https://kbtcoe.org/first-year-engineering-admission/
67743	0	https://kbtcoe.org/admission-me-master-of-engineering/
65405	0	https://kbtcoe.org/about-mvps/
63462	0	https://kbtcoe.org/dseadmission/
63258	0	https://kbtcoe.org/first-year-mba-admission/
62478	0	https://kbtcoe.org/department-of-information-technology/
62196	0	https://kbtcoe.org/masters-of-business-administration-2/
61025	0	https://kbtcoe.org/computer-department/
60414	0	https://kbtcoe.org/masters-of-mechanical-engineering/
58861	0	https://kbtcoe.org/electronics-and-telecommunication-department/
58706	0	https://kbtcoe.org/instrumentation-and-control-department/
58672	0	https://kbtcoe.org/mechanical-department/
58268	0	https://kbtcoe.org/civil-department/
58175	0	https://kbtcoe.org/college-at-a-glance-2/
57910	0	https://kbtcoe.org/committees/
57636	0	https://kbtcoe.org/about/ceos-desk/
57572	0	https://kbtcoe.org/principals-desk/
56904	0	https://kbtcoe.org/first-year-engineering/
56772	0	https://kbtcoe.org/masters-of-civil-engineering/
56524	0	https://kbtcoe.org/vice-principals-desk/
56392	0	https://kbtcoe.org/sarchitnis-message/
55651	0	https://kbtcoe.org/vision-and-mission/

Fig. 2.1

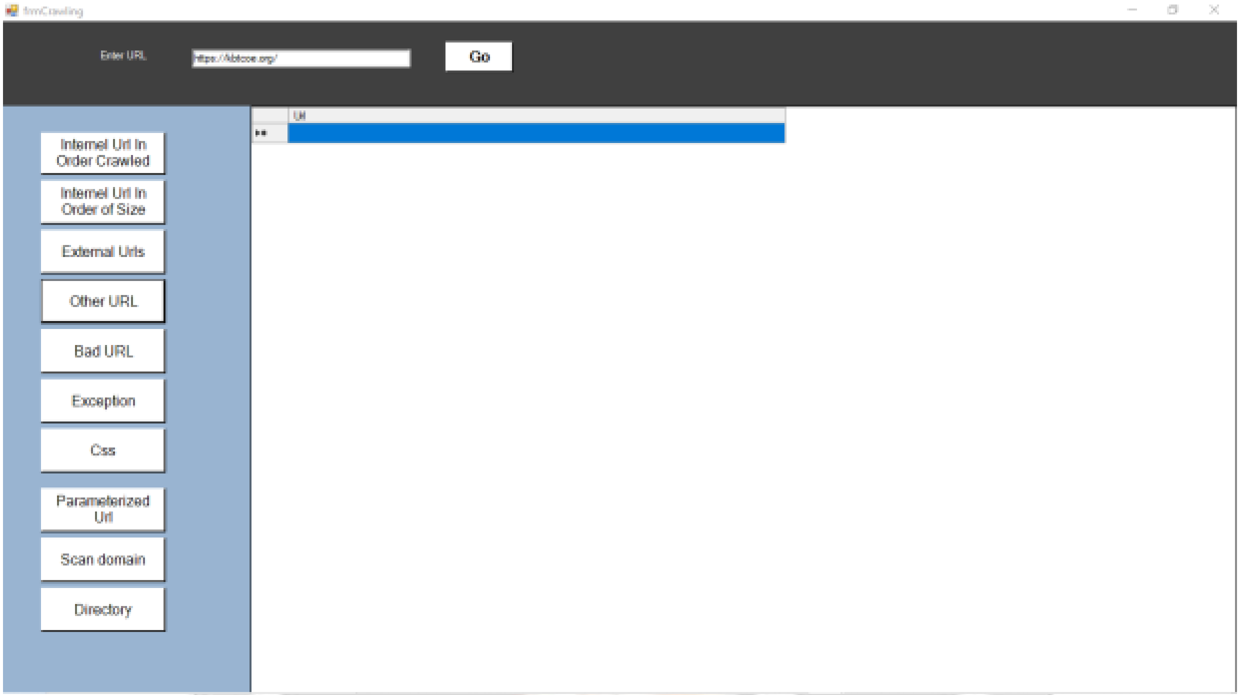
Next tab is External URL in this tab the external URL's are shown. External URL such as the other URL's apart from that website. This is shown in fig 3.1



https://kbtcoe.org/

Url
https://www.facebook.com/mvpkbtcollege
https://www.instagram.com/mvpkbtcoe/
http://www.kbtcoe.org
http://117.239.47.101/login/index.php
http://kbtcoe.org/wp-content/uploads/2019/07/R-D-cell-committee.pdf
http://kbtcoe.org/wp-content/uploads/2019/05/Educational-Verification.pdf
http://alumni.kbtcoe.org
http://www.kbtcoealumni.org
http://kbtcoe.org/congratulations-to-placed-students-2/
http://erp.kbtcoe.org
http://kbtcoe.org/t-p/congratulations-to-placed-students-in-infosys
https://docs.google.com/forms/d/[illegible]
https://docs.google.com/forms/d/[illegible]
http://10.85.0.11/Student%20Log/OPAC.aspx
http://www.[illegible]-courts.com
https://goo.gl/maps/[illegible]
https://docs.google.com/forms/d/e/1FAIpQL[illegible]
https://docs.google.com/forms/d/e/1FAIpQLScw[illegible]
https://docs.google.com/forms/d/e/1FAIpQLSep_[illegible]
https://docs.google.com/forms/d/e/1FAIpQLSd[illegible]/view...
https://[illegible].edu.in/
http://kbtcoe.org/wp-content/uploads/2016/12/Governing-body-council.pdf
http://kbtcoe.org/wp-content/uploads/2016/12/local-managment-committe.pdf
http://kbtcoe.org/wp-content/uploads/2017/10/EXECUTIVE-COUNCIL.pdf
http://kbtcoe.org/wp-content/uploads/2016/12/Revised-International-Advisory-Board.pdf
http://kbtcoe.org/wp-content/uploads/2016/12/Revised-National-Advisory-Board.pdf
http://kbtcoe.org/wp-content/uploads/2016/12/Revised-Local-Advisory-Board.pdf
http://kbtcoe.org/wp-content/uploads/2016/12/Deans-in-various-section.pdf
http://kbtcoe.org/wp-content/uploads/2016/12/IQAC-Composition.pdf

Fig. 3.1

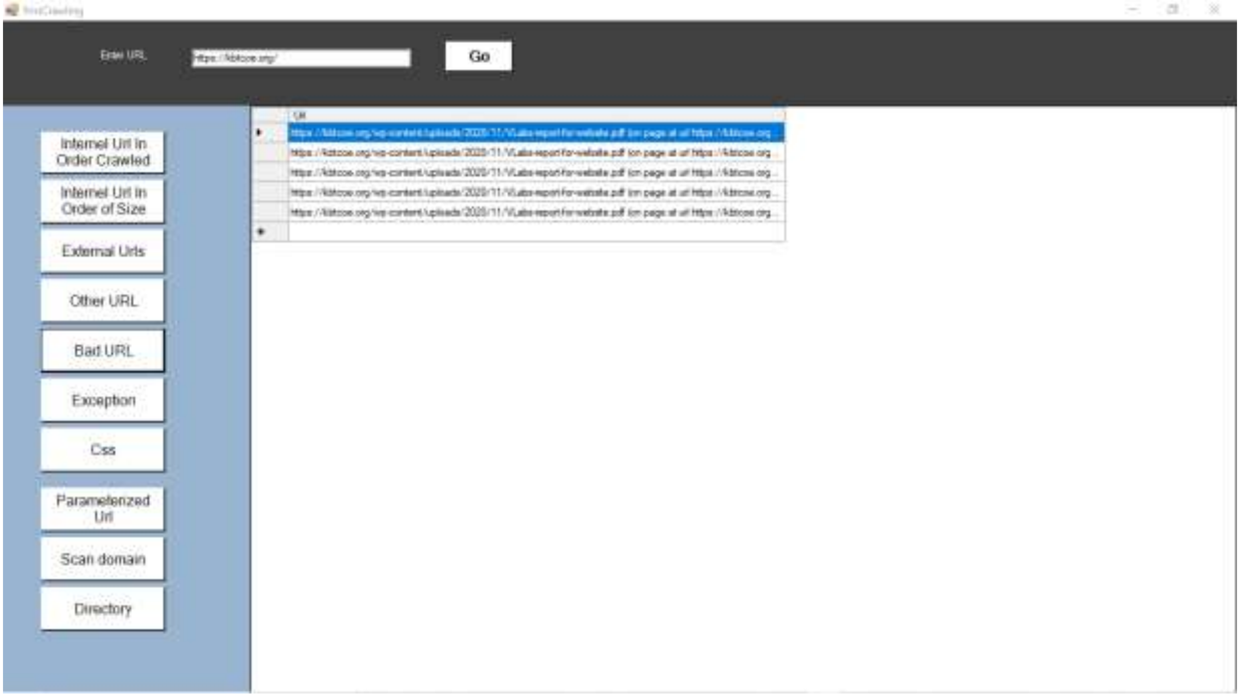


Enter URL https://kbtcoe.org/

Url

Fig. 4.1

Above image is shows the other tab in other tab shows the other type of URL's present in website. This is shown in fig 4.1.



Enter URL https://kbtcoe.org/

Url
https://kbtcoe.org/wp-content/uploads/2020/11/VLabs-report-for-website.pdf (on page at url https://kbtcoe.org...
https://kbtcoe.org/wp-content/uploads/2020/11/VLabs-report-for-website.pdf (on page at url https://kbtcoe.org...
https://kbtcoe.org/wp-content/uploads/2020/11/VLabs-report-for-website.pdf (on page at url https://kbtcoe.org...
https://kbtcoe.org/wp-content/uploads/2020/11/VLabs-report-for-website.pdf (on page at url https://kbtcoe.org...
https://kbtcoe.org/wp-content/uploads/2020/11/VLabs-report-for-website.pdf (on page at url https://kbtcoe.org...

Fig 5.1

Above image shows the Bad URL tab. In Bad URL tab it display's the bad URL present in the website. Such as incorrect format of URL or inaccessible by the application.

Next tab is a Exception tab. It shows any kind of exception is occur in the process of crawling of website. This tab is shown in below in fig 6.1

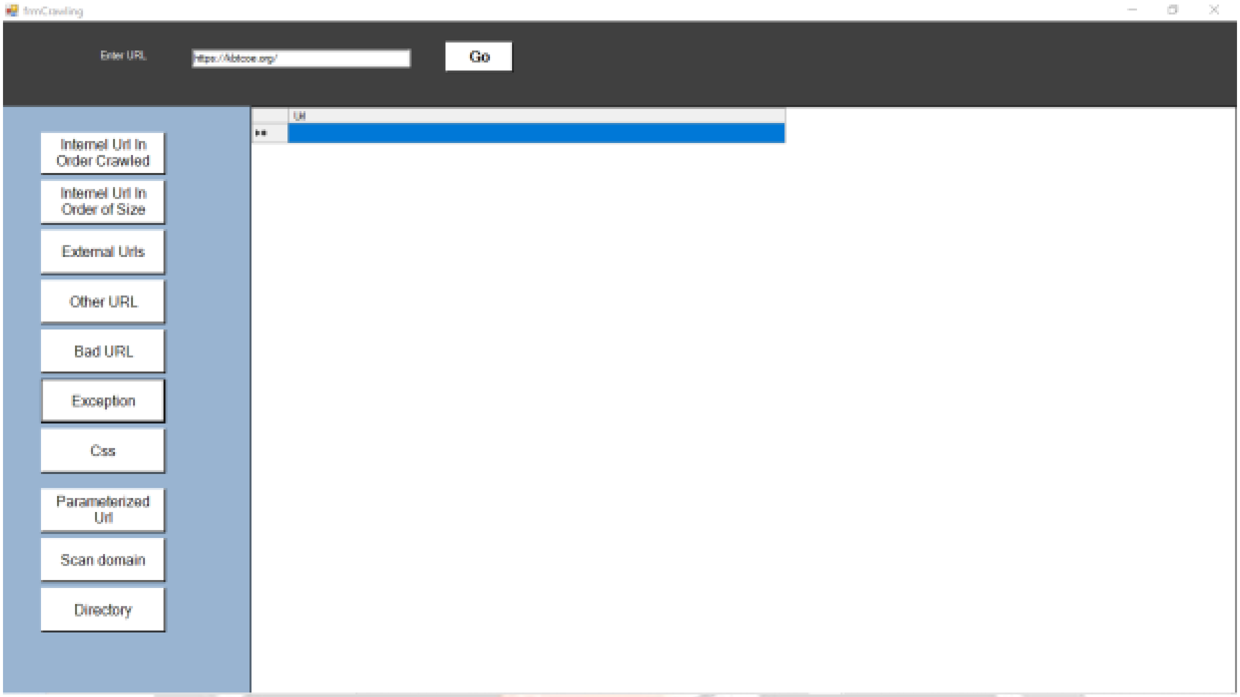
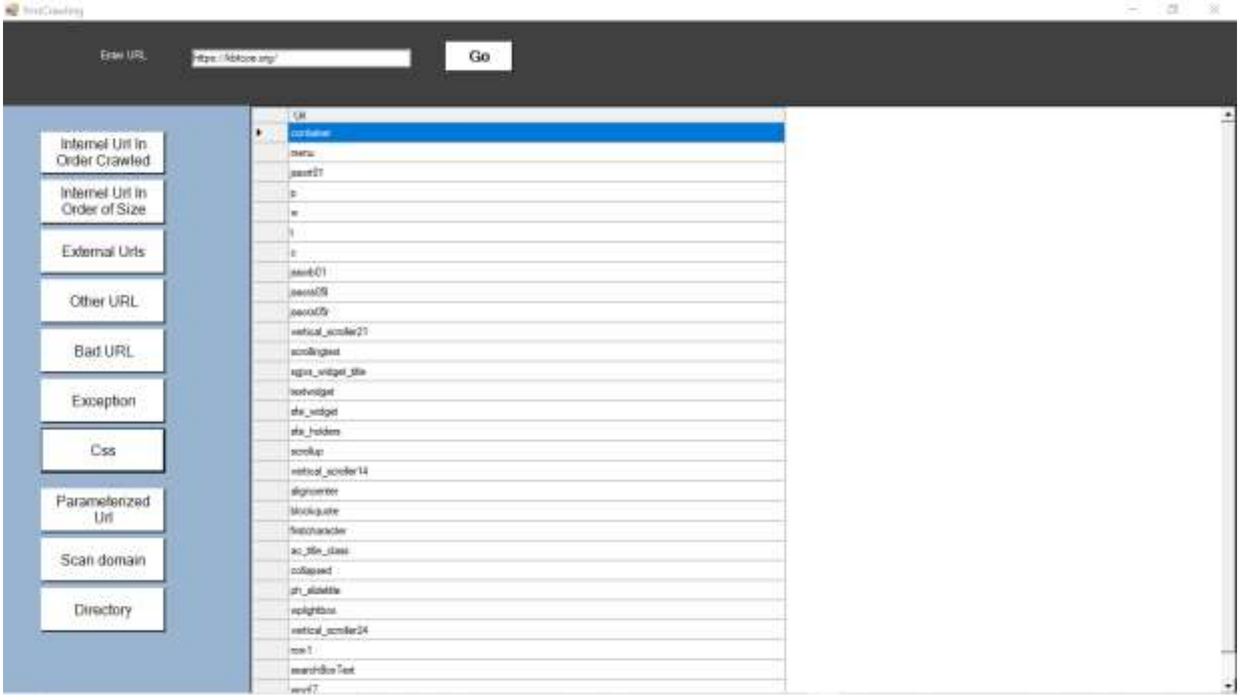


Fig 6.1

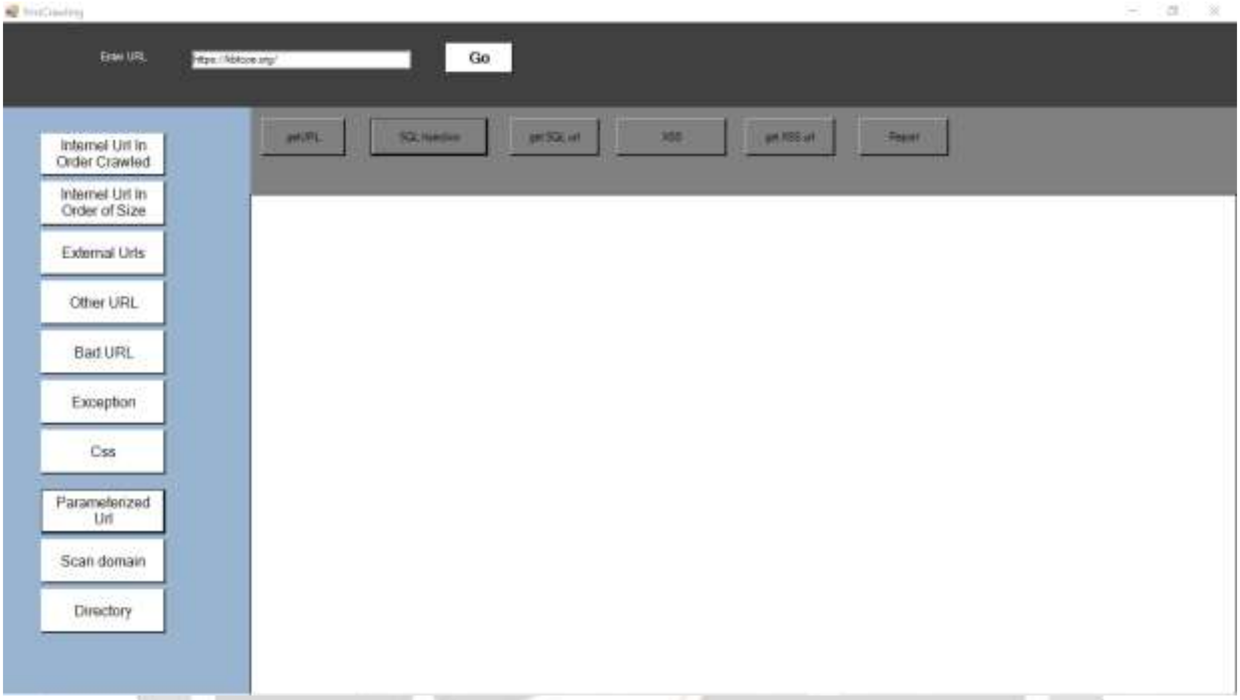
Next tab is CSS tab. In CSS tab we shows the different kind of styles used in the webpage. This tab is shown in below in fig 7.1



Url
container
menu
pasot21
p
w
t
c
pasrb01
paoo09
paoo05r
vertical_scroller21
scrollingtext
sgpx_widget_title
textwidget
sfe_widget
sfe_holders
scrollup
vertical_scroller14
aligncenter
blockquote
firstcharacter
ac_title_class
collapsed
ph_slidetitle
wplightbox
vertical_scroller24
row1
searchBoxText
wevt7

Fig 7.1

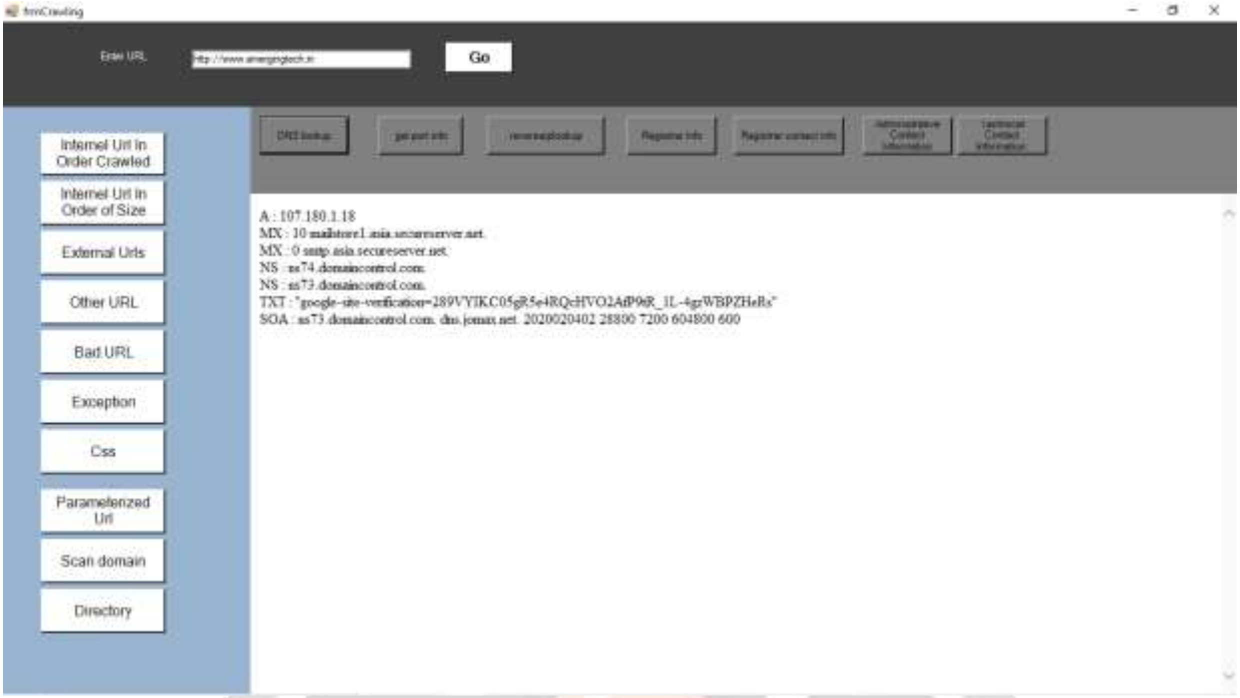
Next tab is Parameterized URL in this tab we shows the URL’s where parameters are passed. And also we perform SQL Injection attack and Cross Site Scripting attack. While performing these attacks we check the website is secure or not. The attack is successfully perform, means the website is not secure for users. This tab shown below in fig 8.1



Enter URL https://kbtcoe.org/

Fig 8.1

Next tab is Scan Domain. In Scan Domain tab we add some more tabs such dns lookup, get port info, reverse ip lookup, Registrar Info, Registrar Contact Info, Administrative Contact Info, Technical Contact Info. The dns lookup tab shows where the website is hosted. Next tab get port info shows how many ports are open or closed. Next tab Reverse Lookup shows the list of websites hosted, where I hosted my website. Next tab is Registrar Info in this tab is shows when the website was register, when is expires, when it is updated. Next tab Registrar contact info shows the name of owner of the website. Next tab Administrative contact Information it shows the office contact information. Next is Technical Contact Information it shows the information about Technical staff. This is shows in fig 9.1



```
A : 107.180.1.18
MX : 10 mailstore1.asia.secureserver.net.
MX : 0 smtp.asia.secureserver.net.
NS : ns74.domaincontrol.com.
NS : ns73.domaincontrol.com.
TXT : "google-site-verification=289VYIKC05gR5e4RQcHVO2AfP9tR_1L-4grWBPZHeRs"
SOA : ns73.domaincontrol.com. dns.jomax.net. 2020020402 28800 7200 604800 600
```

Fig 9.1

Next and last tab is Directory is shows the pdf files, doc files that can be downloaded. Using this tab we can download the files present in website. This is shown in below in 10.1

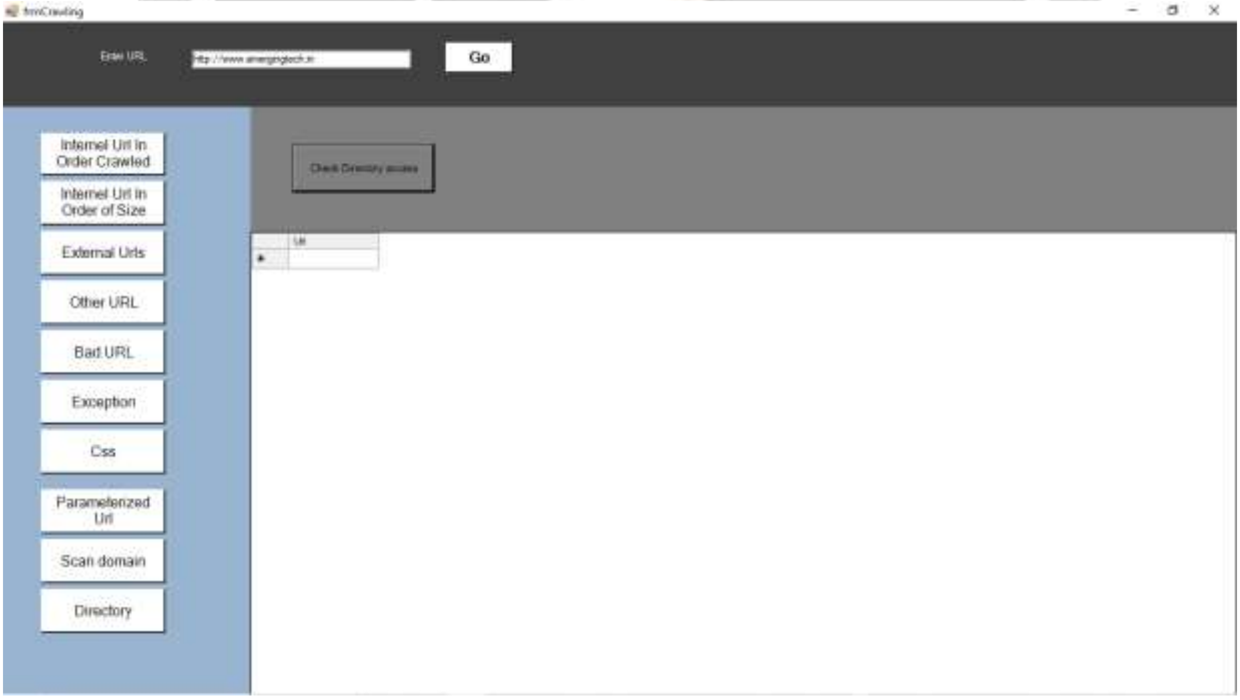


Fig 10.1

VI. CONCLUSION

In the last few years malware have become a significant threat. what are the machine learning algorithms they used in their work, from what sources dataset is collected, what are parameters they consider to reach their goal and the corresponding experimental results In the discussion, it clearly identifies that machine learning algorithms are very useful for the classification and clustering of malware samples for small datasets and for large volumes of data. The worm malware can use the computer ports as gateways to access the computer and invade the network. Our networks need protection from outside attackers to defy against their attacks, so strong systems are needed to detect and prevent the malware from breaking through the networks and do its malicious work.

ACKNOWLEDGMENT

We extend our sincere thanks to Ms. J. V. Shinde ma’am for helping us and guiding us.

REFERENCES

- *Gavrilut, M. Cimpoeşu, D. Anton and L. Ciortuz, "Malware detection using machine learning", Computer Science and Information Technology 2009. IMCSIT'09. International Multiconference on, pp. 735-741, October 2009.*
- *K. Chumachenko, Machine Learning Methods for Malware Detection and Classification, 2017.*
- *L. Liu, B. S. Wang, B. Yu and Q. X. Zhong, "Automatic malware classification and new malware detection using machine learning", Frontiers of Information Technology & Electronic Engineering, vol. 18, no. 9, pp. 1336-1347, 2017.*
- *O. E. David and N. S. Netanyahu, "Deepsign: Deep learning for automatic malware signature generation and classification", Neural Networks (IJCNN) 2015 International Joint Conference on, pp. 1-8, July 2015.*
- *P. Singhal and N. Raul, Malware detection module using machine learning algorithms to assist in centralized security in enterprise networks, 2012,*
- *Masabo, K. S. Kaawaase and J. Sansa-Otim, "Big data: deep learning for detecting malware", Proceedings of the 2018 International Conference on Software Engineering in Africa, pp. 20-26, May*
- *T. N. Phyu, "Survey of classification techniques in data mining", Proceedings of the International MultiConference of Engineers and Computer Scientists, vol. 1, pp. 18-20, March 2009*
- *D. Chen, “Detecting Hiding Malicious Website Using Network Traffic Mining Approach,” 2010 2nd Int. Conforence Educ. Technol. Comput., 2010*
- *Y. Park, D. S. Reeves, and M. Stamp, “Deriving common malware behavior through graph clustering,” Comput. Secur., vol. 39, pp. 419–430,*
-