Detection of Malicious Web Contents using Machine and Deep Learning Approaches

[1]Aasha Singh, [2]Dr. Awadhesh Kumar, [3]Dr. Ajay Kumar Bharti, [4]Dr. Vaishali Singh

[1]Research Scholar, MUIT, Lucknow
[2]KNIT Sultanpur
[3]BBD University, Lucknow
[4]MUIT, Lucknow

ABSTRACT

Websites have been the main target of intruders due to the fast progression of the Internet. An invader implants malicious content in a website page in order to perform a variety of bad and unwanted actions, such as stealing credentials and resources, tempting a web handler to an unsafe website, installing or downloading software to link a botnet, or participating in dispersed denial of service attacks. It can also damage user's system. Uninvited web content such as phishing, spam, and drive-by-downloads are hosted on malicious URLs, which entice unsuspecting users to become victims of schemes such as financial loss, data theft, and malware installation. Every year, billions of dollars are lost as a result of this. It is critical to detect and respond to such dangers as soon as possible.

Keywords: Web content, URL, Cyber-crime, malware, Classification.

1. INTRODUCTION

Nowadays after covid, there is a very heavy usage of internet, either in the form of distance learning or using it for company team meetings. Using of internet can also cause Cyber-crime using malicious URL that they can fetch, read and manipulate user data [11]. So, it's very important to know that the page user is visiting is safe for them of not. Under this assignment work we have tried several approaches for designing a model that helps us to determine the category of the URL that the user is visiting. As the quantity of web pages grows, so do the number of rouge web pages, and the attack becomes more sophisticated [12]. The aim of malicious URL identification is to preclude the company employees from log on websites that may affect with the maneuver of the business – such as websites that are not associated to the work, websites with distasteful or unlawful Web Content, or websites related with phishing efforts. Whereas unrestricted website surfing and accessing might be very beneficial for the employees and can create them all much more productive, this can also uncover administrations to a extensive variety of security threats, such as dissemination of intimidations, data loss or removal, or legal issues. The web data presentation has become a primary objective for cyber offenders by inoculating malware specifically JavaScript to accomplish malevolent actions for impersonation [10]. Thus, it becomes an imperious to discover such malevolent code in real time before any spiteful action is performed. We present an analysis for detecting a malicious web page using machine and deep learning approaches here. The analysis of results shows that other techniques can competently classify spiteful code from benign code with promising result.

2. LITERATURE REVIEW

- S Sananse and Sarode in 2015, developed a technique for detecting phishing and non- phishing. In their research paper they have used Random Forest and Content-based algorithm for classification on the dataset.

- Jeeva and Raj Singh in 2016, classify important characteristic's that distinguish between benign and phishing URLs. In their research they have used mining association rules for features to detect phishing URLs. But they only focused on two categories of URLs i.e. benign and phishing.

- Patil and Patil in 2016, utilized the URL string's static analysis to detect malicious web pages effectively. They used 79 static features of URLs that extracted from characteristics of benign and malicious URLs. They assessed machine learning algorithms on their dataset and their experimental analysis have showed a detection rate between 95% - 99%.

- Some scientists suggested URLNet, which is a convolutional neural network (CNN) technique for detection of malicious URLs using deep neural networks. Their model instigated unadventurous CNN for all words and characters of the URL (Le, Pham, Sahoo, and Hoi, 2018).

- An analysis and study proposed in this paper propose a neural network based approach; in this research work, they have used deep learning with convolutional neural network methodology that detects structures such as malicious URLs, files, and registry keys (Saxe and Berlin, 2017).

- A latest study fortifying URL address, a method using Event De-noising Convolutional Neural Network for Sequence Detection in malicious URL. Here the authors proposed a model to detect series of malicious URL from proxy logs with a low false positive rate (Shibahara - 2017).

- EDCNN is a specific CNN to decrease the negative impact of benign URLs redirected from compromised websites.

- Vazha yil, Vinaya kumar and Soman (2018), presented a comparative study between classical machine learning techniques and deep learning methods to detect malicious URLs.

3. PROBLEM STATEMENT AND DATASET

To detect malicious URL categorically, we have used ISCX-URL-2016 dataset to evaluate accuracy of model. There are five types of malicious URL that are Benign, Spam, Phishing, Malware and Defacement that are mentioned in dataset and our main goal is to get better accuracy in classification of dataset by making minor changes in their approach or algorithm. We have used 70% dataset for training and rest 30% of the dataset used for testing purpose.

3.1 Description of Dataset

We have taken divided dataset into two parts:

(a) Training dataset: For training dataset, we have different categories of dataset like: Benign, Spam, Phishing, Malware, and Defacement.

(b) Testing dataset: Same categories of dataset used for testing purpose.

Data	Benign	Spam	Phishing	Malware	Defacement	Total
Training	5478	4662	5286	4696	5565	25687
Testing	2303	2036	2291	2015	2365	11010

Table 1: Dataset used for detection of malicious web contents

3.2 Modified Approach

- **Data pre-processing:**

For the pre-processing of data, first we sliced the URLs to use it as a features by “/” , “-”, “.” and “com”.

- **For Example: http://1337x.to/torrent/1110018/Blackhat-2015-...** is changed into**:**

⇩

['', 'to', 'torrent', '720p', 'h264', 'dl', 'dd5', "b'http:", "'", '2015', 'russian', 'rufgt', '1', 'web', '1337x.to', 'blackhat', '1110018', '1337x']

- **Feature Extraction:** Then we have used “**TfidfVectorizer”** for extraction of feature from text words. We have used 70% data for training and rest 30% data for testing.

3.3 Machine Learning Approaches

Machine Learning techniques offer a system that accomplishes (1) to learning by itself (2) and provide advancement from past experience (3) it does all without devising any definite program. Machine learning has the main ability to deliver capability to the computer system to acquire knowledge by design without any intervention of human being. Here we are applying two machine learning approaches for classification of data: (i) SVM and (ii) Random Forest

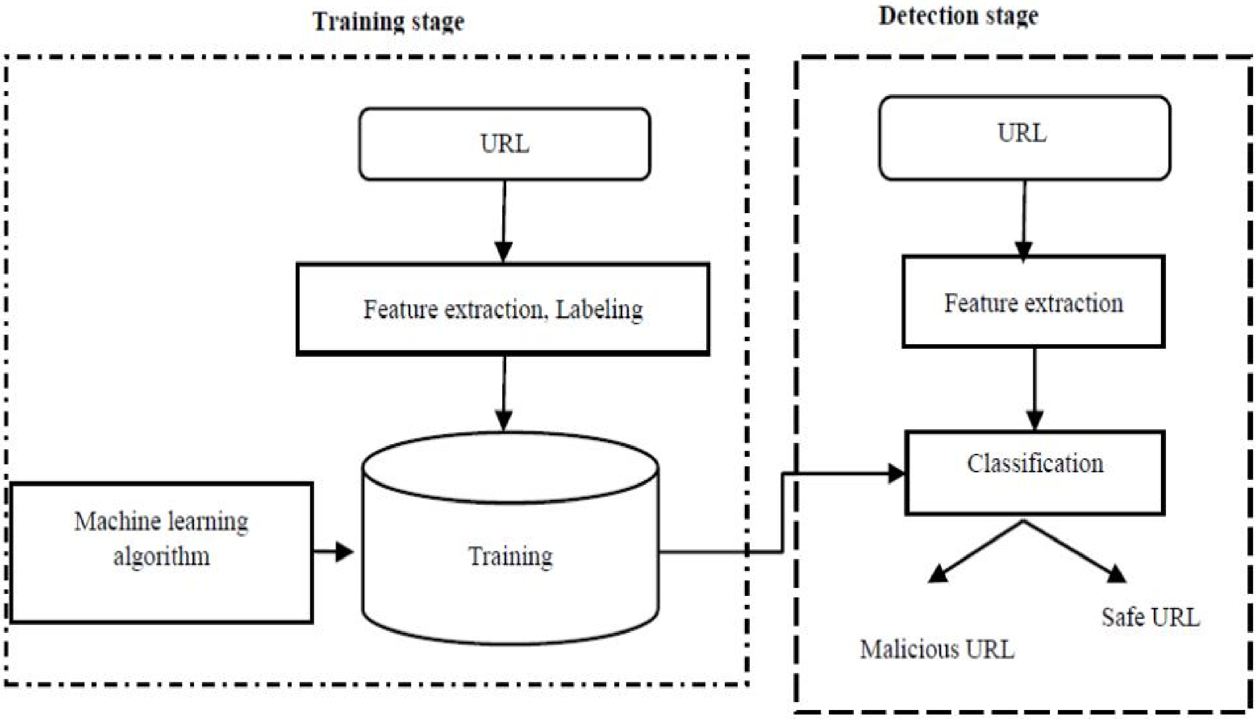
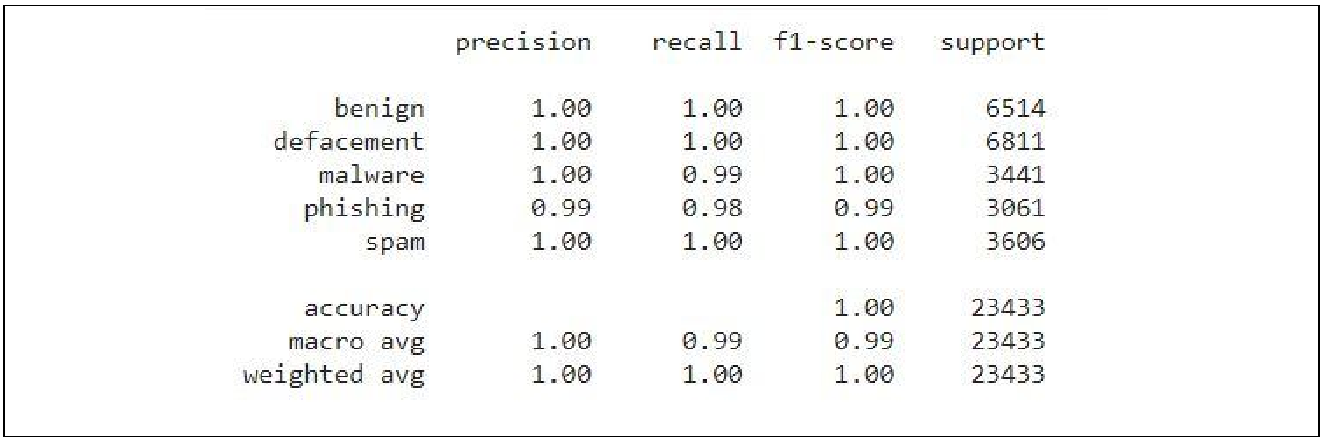


Fig. 1. Malicious URL Detection Model using Machine Learning.

3.3.1 SVM

We have used Support Vector Classification from SVM to classify different malware URL classes. We have taken penalty value C=2 i.e., if a class belongs to wrong hyper plane then this will cost penalty.



```
              precision    recall  f1-score   support

      benign       1.00      1.00      1.00      6514
 defacement       1.00      1.00      1.00      6811
     malware       1.00      0.99      1.00      3441
    phishing       0.99      0.98      0.99      3061
        spam       1.00      1.00      1.00      3606

    accuracy                           1.00     23433
   macro avg       1.00      0.99      0.99     23433
weighted avg       1.00      1.00      1.00     23433
```

Fig. 2: Classification Report of SVM

3.3.2 Random Forest

We have taken 10 trees with random state 30 in our classification. There are 10 trees with random rows and features values and make decisions on class to learn.

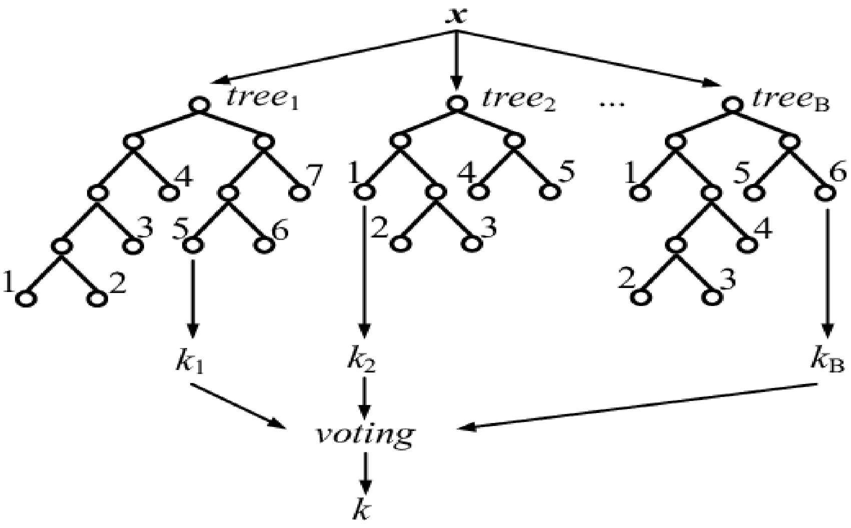


Fig. 3: Random Forest

4. DEEP LEARNING APPROACH

There are so many other methods and features designed by various experts that provide improvements in the performance for prediction models [1]. These procedures undergo with a number of limitations: (i) Incapability to effectually detention semantic implication and consecutive configurations in URL strings; (ii) Demanding considerable handbook feature engineering; and (iii) Incapability to handle unnoticed structures and simplify to check dataset. Deep Learning technique relates to the machine learning as it is a form of this technique but diverges in the usage of Neural Networks where we rouse the task of a brain to a definite amount and use a 3D hierarchy in dataset to recognize patterns that are much more beneficial.

4.1 CNN and LSTM method

According to research paper by Emine Uçar[13], they have used CNN and LSTM model and they got 96% to 98% accuracy. In their research work they have used Layered model. In our research work, we have performed our analysis by implementing the different approach by applying the MLP (Multi-layer Perceptron).

- MLP Classifier stands for Multi-layer Perceptron classifier that it shows the connection by itself to a Neural Network. As the performance of different classification algorithms such as Naive Bayes or Support Vectors Classifier, the MLP.

- Classifier depends on a fundamental Neural Network to accomplish the assignment of classification. In this work, we have used Google colab environmental setting for running code. It gives all the resources that are required to run heavy codes.

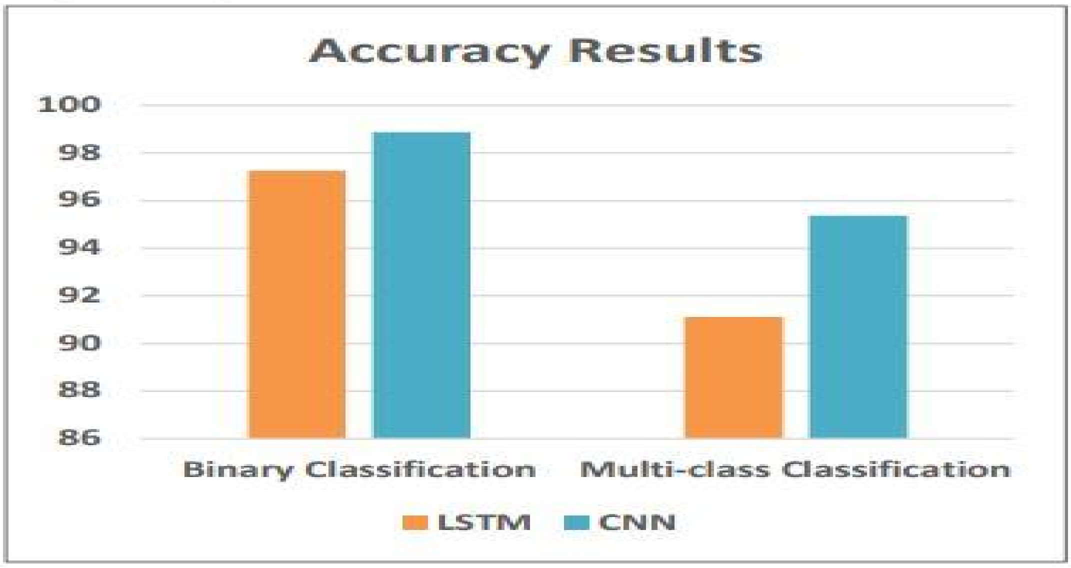


Fig. 4: Accuracy analysis of LSTM and CNN

5. CONCLUSION

In this research paper, we have analyzed the accuracy performance by applying different machine learning techniques and deep learning method. By using SVM (SVC) and Random Forest we get accuracy around 96%. By research paper by Emine Uçar [13], they get accuracy around 96% to 98% by using CNN and LSTM. We have used MLP i.e. Multi-layer perceptron that gives accuracy result between 97% to 99% accuracy.

References

[1] M. Antonakakis, R. Perdisci, D. Dagon, W. Lee, and N. Feamster, “Building a dynamic reputation system for dns,” in Proceedings of the 19th USENIX conference on Security, ser. USENIX Security’10. Berkeley, CA, USA: USENIX Association, 2010, pp. 18–18.

[2] M. Akiyama, M. Iwamura, Y. Kawakoya, K. Aoki, and M. Itoh, “Design and implementation of high interaction client honeypot for drive-by-download attacks,” IEICE Transactions on Communications, vol. E93.B, no. 5, pp. 1131– 1139, 2010.

[3] M. Felegyhazi, C. Kreibich, and V. Paxson, “On the potential of proactive domain blacklisting,” in Proceedings of the 3rd USENIX conference on Large-scale exploits and emergent threats: botnets, spyware, worms, and more, ser. LEET’10. Berkeley, CA, USA: USENIX Association, 2010, pp. 6–6.

[4] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Beyond blacklists: learning to detect malicious web sites from suspicious URLs,” in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ser. KDD ’09. New York, NY, USA: ACM, 2009, pp. 1245–1254.

[5] A. Renjan, K. P. Joshi, S. N. Narayanan and A. Joshi, "DAbR: Dynamic attribute-based reputation scoring for malicious IP address detection", IEEE International Conference on Intelligence and Security Informatics (ISI), pp. 64-69, Nov 2018.

[6] V. Paxson, “Bro: a system for detecting network intruders in real-time,” in Proceedings of the 7th conference on USENIX Security Symposium - Volume 7. Berkeley, CA, USA: USENIX Association, 1998, pp. 3–3.

[7] M. Roesch, “Snort - lightweight intrusion detection for networks,” in Proceedings of the 13th USENIX conference on System administration, ser. LISA ’99. Berkeley, CA, USA: USENIX Association, 1999, pp. 229–238.

[8] M. Ishida, H. Takakura, and Y. Okabe, “High-performance intrusion detection using OptiGrid clustering and grid-based labelling,” in Applications and the Internet (SAINT), 2011 IEEE/IPSJ 11th International Symposium on, Jul. 2011, pp. 11 –19.

[9] C. Seifert, I. Welch, P. Komisarczuk et al., “HoneyC - The low interaction client honeypot,” proceedings of the 2007 NZCSRCS, Waikato University, Hamilton, New Zealand, 2007.

[10] Capture-hpc. [Online]. Available: https://projects.honeynet. org/capture-hpc/

[11] https://www.kaggle.com/cheedcheed/top1m/metadata

[12] http://lists.blocklist.de/lists/all.txt

[13] Emine Uçar, Murat Uçar, Mürsel Ozan İNCETAŞ “A deep learning approach for detection of malicious URLs”; International Management Information Systems Conference “Connectedness and Cyber security” 09-12 October 2019.

[14] Hung; Pham, Hong Quang; Sahoo, Doyen; And Hoi, Steven C. H.. Urlnet: Learning A Url Representation With Deep Learning For Malicious Url Detection. (2017). Pods 2017: Proceedings of the Acm Symposium on Principles of Distributed Computing, Washington, Dc, July 25-27. 1-13. Research Collection School Of Information Systems.

[15] Lekshmi A R1, Seena Thomas2: Detecting Malicious Urls Using Machine Learning Techniques: A Comparative Literature Review. International Research Journal Of Engineering And Technology (Irjet) E-Issn: 2395-0056, Volume: 06 Issue: 06 | June 2019

[16] Doyen Sahoo, Chenghao Liu, Steven C.H. Hoi: Malicious Url Detection Using Machine Learning: A Survey. © 2019 Association For Computing Machinery, Vol. 1, No. 1, Article . Publication Date: August 2019.

[17] Immadisetti Naga Venkata Durga Naveen, Manamohana K, Rohit Verma : Detection Of Malicious Urls Using Machine Learning Techniques: International Journal Of Innovative Technology And Exploring Engineering (Ijitee) Issn: 2278-3075, Volume-8 Issue-4s2 March, 2019.

[18] Tie Li A, Gang Kou B, Yi Peng A,∗: Improving Malicious Urls Detection Via Feature Engineering: Linear And Nonlinear Space Transformation Methods. Information Systems 91 (2020) 101494, Https://Doi.Org/10.1016/J.Is.2020.101494, Published By Elsevier Ltd.

[19] Inayakumar R, Sriram S, Soman Kp, And Mamoun Alazab: Malicious Url Detection Using Deep Learning: Department Of Corporate And Information-Services, Northern Territory