IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 4, April 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304362 |

Analyzing Machine Learning Approaches for Phishing Website Detection

Prof. Monisha R, Megha G, Nayana PGR, Prarthana KM, Sahana CS

Assistant Professor, Department of Computer Science & Engineering, S J M Institute of Technology, Chitradurga, Karnataka, India

Student, Department of Computer Science & Engineering, S J M Institute of Technology, Chitradurga, Karnataka, India

Student, Department of Computer Science & Engineering, S J M Institute of Technology, Chitradurga, Karnataka, India

Student, Department of Computer Science & Engineering, S J M Institute of Technology, Chitradurga, Karnataka, India

Student, Department of Computer Science & Engineering, S J M Institute of Technology, Chitradurga, Karnataka, India

ABSTRACT: With the increasing use of mobile devices, there is a growing trend to move almost all real-world operations to the cyber world. Although this makes easy our daily lives, it also brings many security breaches due to the anonymous structure of the Internet. Used antivirus programs and firewall systems can prevent most of the attacks. However, experienced attackers target on the weakness of the computer users by trying to phish them with bogus webpages. These pages imitate some popular banking, social media, e-commerce, etc. sites to steal some sensitive information such as, user-ids, passwords, bank account, credit card numbers, etc. Phishing detection is a challenging problem, and many different solutions are proposed in the market as a blacklist, rule-based detection, anomaly-based detection, etc. Phishing attacks are malicious attempts to deceive users into revealing sensitive information, such as login credentials or financial details, by masquerading as legitimate entities. These attacks often involve fraudulent websites or emails designed to trick users into disclosing confidential information.

KEYWORDS: Phishing detection, machine learning, deep learning, RNN-GRU, web browser extension.

I. INTRODUCTION

Phishing attacks have become a paramount concern for cybersecurity researchers due to their sophisticated methods of deception, especially in crafting fake websites that closely mimic legitimate ones. This mimicry poses a significant challenge as even knowledgeable users can struggle to distinguish between authentic and fraudulent sites, making them susceptible to falling prey to phishing attacks. The primary objective of these attacks typically revolves around illicitly acquiring sensitive information, with a particular focus on banking credentials, leading to substantial financial losses. For instance, businesses in the United States alone report losses of approximately $2 billion annually due to phishing attacks, as indicated in [1]. The global impact, as estimated in the 3rd Microsoft Computing Safer Index Report of February 2014, could potentially exceed $5 billion [2]. The success of phishing attacks is often attributed to the lack of user awareness regarding these deceptive practices. Since phishing attacks capitalize on exploiting user vulnerabilities, effectively mitigating them proves to be a complex challenge. Nonetheless, enhancing phishing detection techniques is imperative to combat this pervasive threat. The traditional approach to identifying phishing websites involves updating blacklisted URLs and Internet Protocol (IP) addresses in antivirus databases, commonly referred to as the "blacklist" method. However, attackers constantly innovate by employing creative techniques such as URL obfuscation, fast-flux techniques with automated proxies, algorithmic generation of new URLs, and more, which can effectively evade blacklists.

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304362 |

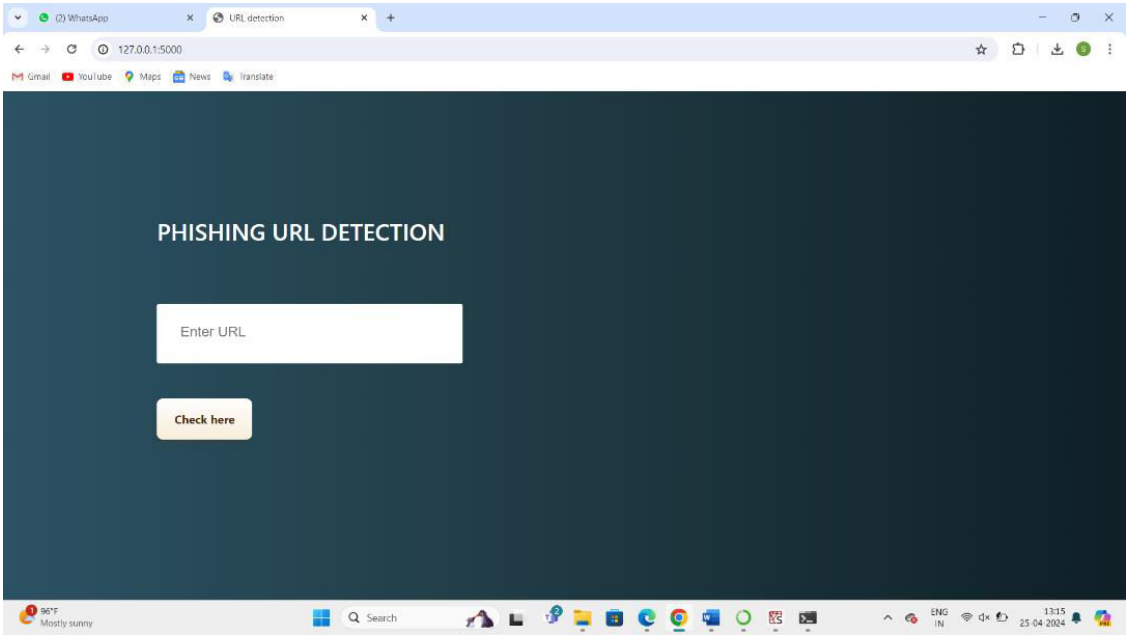
II. RELATED WORK

Phishing attacks represent a serious problem, and the technology for detecting and intercepting phishing attacks is constantly evolving. It is the most accurate and fast way to filter good URLs through the whitelist and block phishing URLs through the blacklist. However, the list method cannot detect new phishing links, and because of the low cost of creating a phishing URL, the attacker does not rely on using the same phishing link multiple times. Many research reports based on machine learning have been published, and high accuracy results have been obtained in experiments. However, in the actual network environment, there are still many victims of phishing attacks every year, causing economic losses. There is still a certain gap between the experimental data results and the real network security solutions. Therefore, it is very important to study antiphishing solutions in a real-time environment. We divide the related work into two parts: (1) deep learning-based methods for detecting phishing websites (2) frameworks with prototype implementations.

III. METHODOLOGY

Machine learning based different algorithms were run in the experiment. These are: Random Forest (RF), Naive Bayes (NB), XG Boost, Gradient descent. XG Boost is an algorithm based on gradient boosted decision trees that put speed and performance in the foreground. Naive Bayes is a classification algorithm based on conditional probability. It works according to Bayes' theorem. It is preferred due to its easy implementation and less training time. Gradient descent is by far the most popular optimization strategy used in machine learning and deep learning at the moment. Random Forest is an algorithm that works with the Ensemble Learning technique by creating a large number of trees in the dataset. It divides into subtrees

IV. EXPERIMENTAL RESULTS



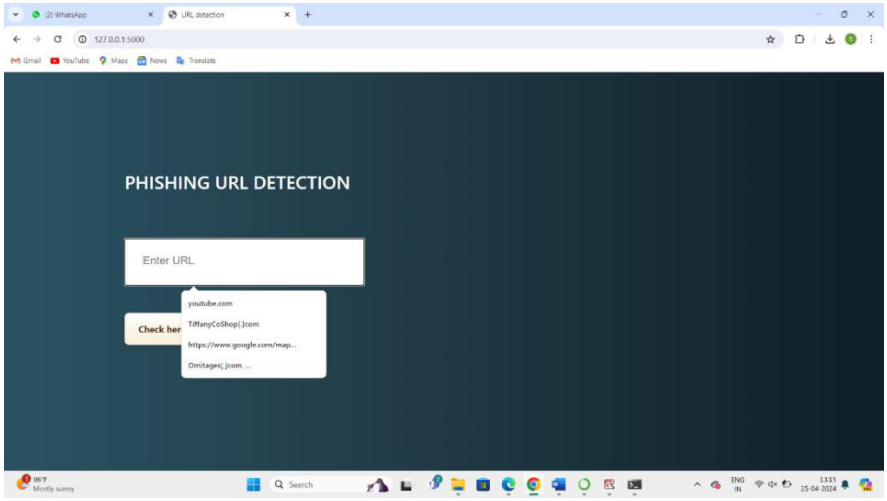
Fig(a): Home page of Phishing URL Detection

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

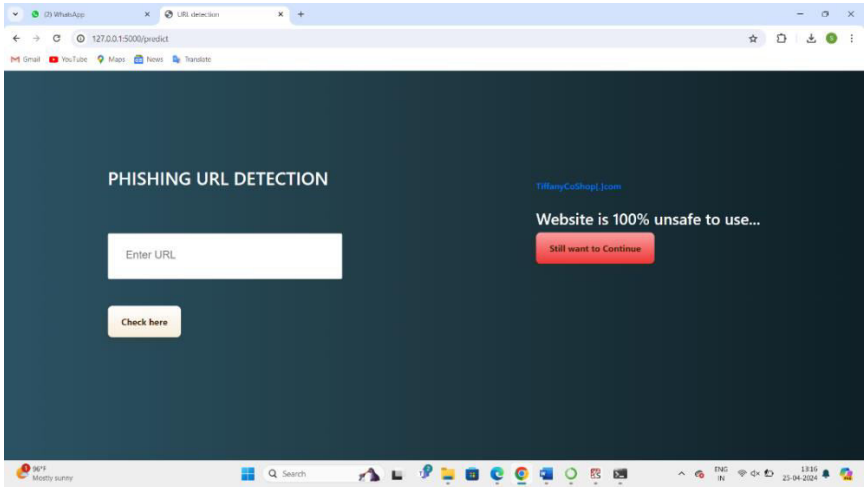
| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

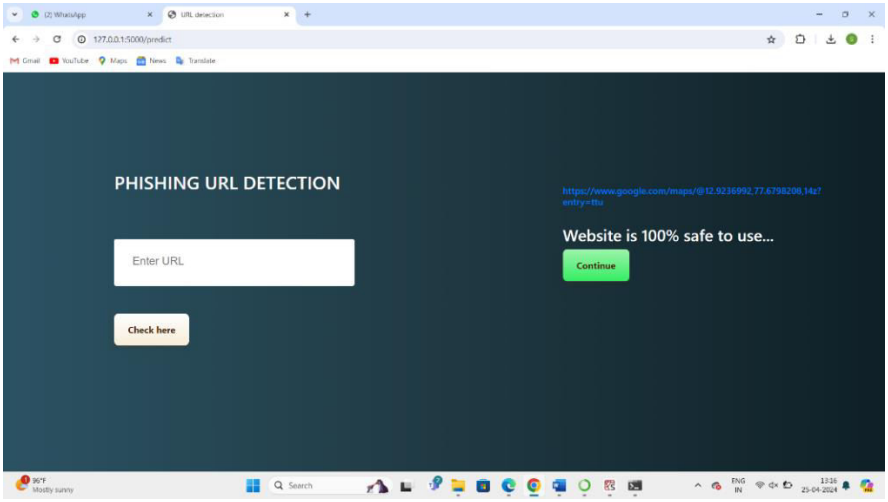
| DOI:10.15680/IJIRSET.2024.1304362 |



Fig(b): Entering the URL of websites



Fig(c): Result showing Website is unsafe to use



Fig(d): Result showing Website is Safe to use

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304362 |

V. CONCLUSION

This project aims to enhance detection method to detect phishing websites using machine learning technology. In this project we are using RFEC method for the feature extraction we achieved 97.14% detection accuracy using gradient boosting algorithm with lowest false positive rate. Also result shows that classifiers give better performance when we used more data as training data. In future hybrid technology will be implemented to detect phishing websites more accurately, for which random forest algorithm of machine learning technology and blacklist method will be used.

REFERENCES

[1] Gunter Ollmann, "The Phishing Guide Understanding & Preventing Phishing Attacks", IBM Internet Security Systems, 2007.
[2] https://resources.infosecinstitute.com/category/enterprise /phishing/the-phishinglandscape/phishing-data-attack statistics/#gref
[3] Mahmoud Khonji, Youssef Iraqi, "Phishing Detection: A Literature Survey IEEE, and Andrew Jones, 2013
[4] Mohammad R., Thabtah F. McCluskey L., (2015) Phishing websites dataset. Available: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites Accessed January 2016
[5] http://dataaspirant.com/2017/01/30/how-decision-treealgorithm-works/
[6] http://dataaspirant.com/2017/05/22/random-forestalgorithm-machine-learing/
[7] https://www.kdnuggets.com/2016/07/support-vectormachines-simple-explanation.html
[8] www.alexa.com
[9] www.phishtank.co

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details