International Journal of Web of Multidisciplinary Studies
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
website: http://ijwos.com
Vol.02 No.11.
OPEN ACCESS
E-ISSN : 3049-2424
DOI:

ANALYZING MACHINE LEARNING APPROACHES FOR PHISHING WEBSITE DETECTION

Ananyashree C M[1], Mrs.Manjula K[2]

[1] *Master in Computer Applications, Faculty of Computing and IT, GM University, Davanagere-577006, Karnataka,India*
[2] *Assistant Professor, Faculty of Computing and IT, GM University, Davanagere-577006, Karnataka, India.*

Article Info

Article History:

Published:11 Nov 2025

Publication Issue:
Volume 2, Issue 11
November-2025

Page Number:
249-254

Corresponding Author:
Ananyashree C M

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, ete. 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.

1.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 S2 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 S5 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.

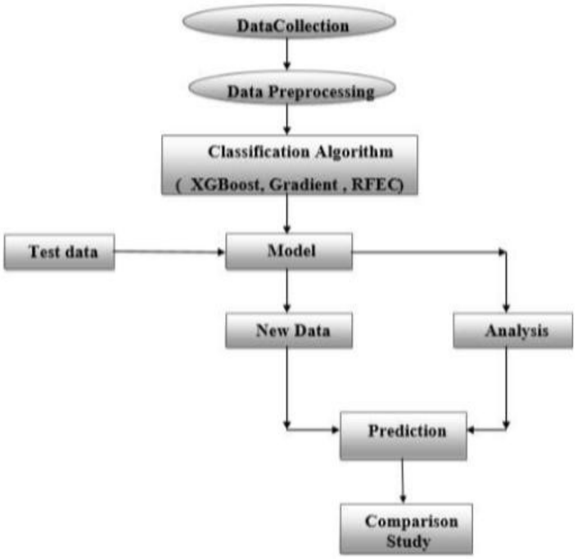
2.RELATED WORK

Phishing atacks represent a serious problem,and the technology for detecting and intereepting phishing attacks is constantly evolving.It is the most accurate and fast way to fiter 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 atiacks 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 withprototype implementations.

3. METHODOLOGY

Machine learming based different algorithms were run in the experiment.These are:Random Forest (RF),Naïve 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.

System Architecture

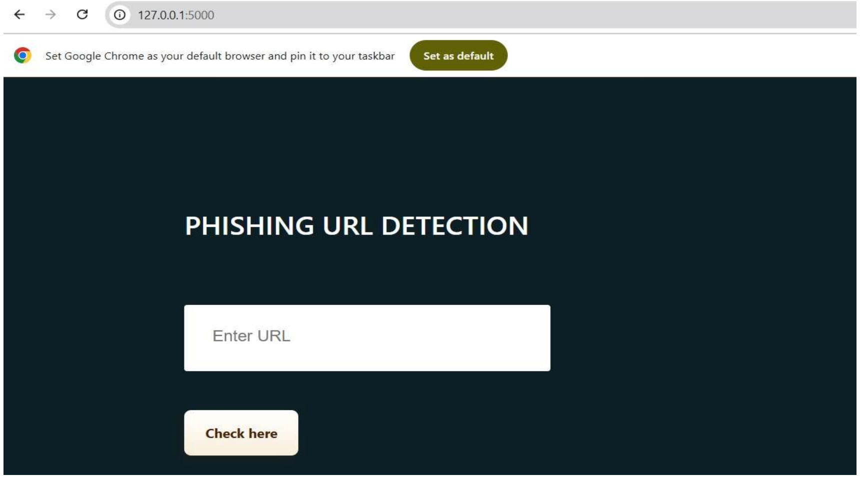


4. PROPOSED SYSTEM

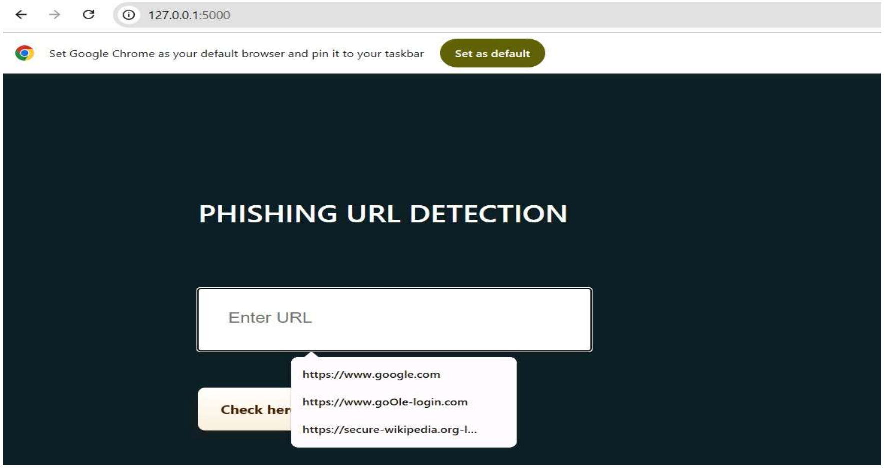
The proposed system uses machine learning algorithms such as XGBoost and Gradient Boosting to detect phishing websites based on URL features. It analyzes characteristics like the presence of IP addresses, HTTPS tokens, special symbols, and SSL certificates to classify websites as legitimate, suspicious, or phishing. Using the Recursive Feature Elimination with Cross-Validation (RFEC) method, the system selects the most relevant features to improve accuracy. With an achieved accuracy of about 95%, the system provides real-time detection through a user interface where users can input a URL and instantly know if it is safe or unsafe.

5.RESULTS AND DISCUSSION

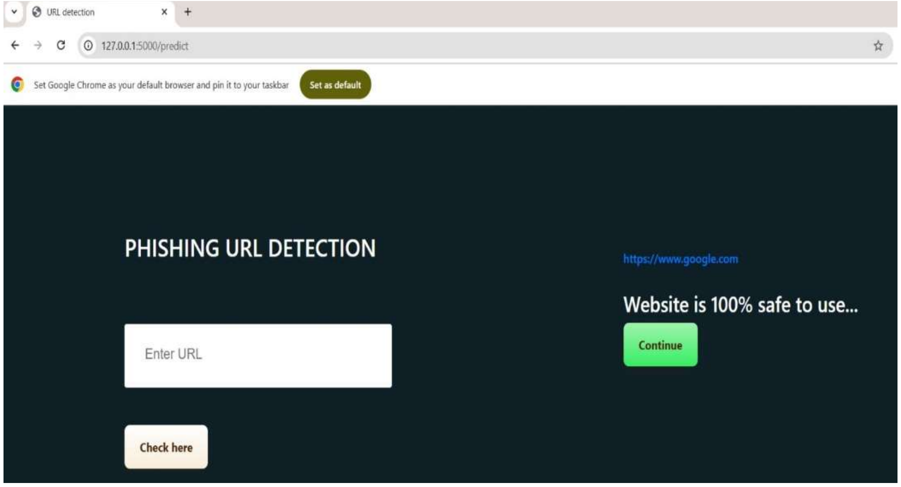
The proposed system for phishing website detection was implemented using Python with Gradient Boosting and XGBoost algorithms, alongside Recursive Feature Elimination with Cross-Validation (RFEC) for effective feature extraction. Using a dataset of 1,353 real-world URLs labeled as legitimate, suspicious, or phishing, the models analyzed features such as the presence of IP addresses, special symbols, URL length, HTTPS tokens, redirection services, and SSL certificate validity. The Gradient Boosting algorithm achieved the highest detection accuracy of approximate 90% with a low false positive rate, while XG Boost achieved 90% above accuracy. The system interface allows users to input a URL and instantly receive results indicating whether the website is safe or unsafe. The findings clearly demonstrate that machine learning-based approaches outperform traditional blacklist or heuristic-based systems by providing adaptability, scalability, and the ability to detect zero-hour phishing attacks. The RFEC technique significantly improved model efficiency by focusing on the most relevant features, thereby reducing noise and computational complexity. Overall, the project highlights that ensemble learning models like Gradient Boosting offer a robust and intelligent defense mechanism against phishing threats by ensuring high accuracy and minimal false positives. In the future, integrating hybrid models such as Random Forest with blacklist verification and expanding datasets to include multilingual domains could further enhance system reliability, while real-time integration through browser extensions or email filters would provide users with proactive phishing protection.



Fig(a): Home page of Phishing URL Detection


127.0.0.1:5000
Set Google Chrome as your default browser and pin it to your taskbar
Set as default
PHISHING URL DETECTION
Enter URL
https://www.google.com
https://www.goOle-login.com
https://secure-wikipedia.org-l...
Check her

Fig(b): Entering the URL of Websites



URL detection

127.0.0.1:5000/predict

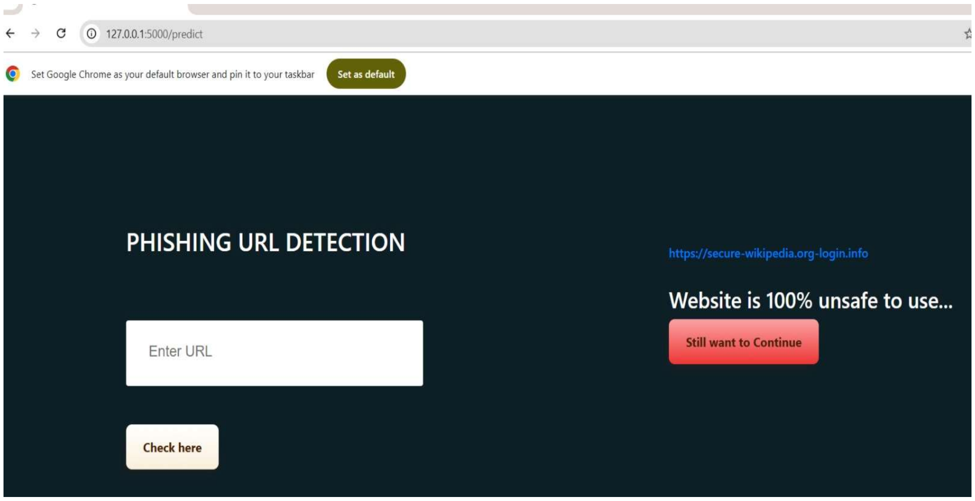
Set Google Chrome as your default browser and pin it to your taskbar

PHISHING URL DETECTION

https://www.google.com

Website is 100% safe to use...

Fig(c): Result showing Websites is Safe to use



127.0.0.1:5000/predict

Set Google Chrome as your default browser and pin it to your taskbar

PHISHING URL DETECTION

https://secure-wikipedia.org-login.info

Website is 100% unsafe to use...

Fig(d): Result showing Websites is Unsafe to use

6. 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 approximate 90% above 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.

7.AKNOWLEDGEMENT

The successful completion of this research paper would not have been possible without the guidance, encouragement, and support of many individuals. First and foremost, I express my deepest gratitude to my research guide **Prof.Manjula K** for their continuous motivation, expert advice, and valuable insights throughout the course of this study. I also extend my sincere thanks to the faculty and staff of **GM UNIVERSITY** for providing the necessary resources and infrastructure to carry out this research. Special thanks to my friends and peers for their constant support and encouragement, which kept me motivated throughout this journey.

Finally, I acknowledge all the researchers and developers whose work in the fields of cloud computing, serverless technology, and cybersecurity has laid the foundation for this study.

References

[1] Gunter Ollmann, “The Phishing Guide Understanding and Preventing Attacks”, IBM Internet Security Systems, 2007.

[2] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, “Machine learning based phishing detection from urls,” Expert Systems with Applications, vol. 117, pp. 345–357, 2019.

[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-/

[6] http://dataaspirant.com/2017/01/30/how-decision-treealgorithm-works/

[7] www.alexa.com

[8] www.phishtank.com