Project Report on Machine learning-based phishing detection from URLs

Abir Mahmud

This report is submitted to the school of Business and Economics,
United International University as a partial requirement for the degree fulfillment of
Bachelor of Business Administration

Machine learning-based phishing detection from URLs

Submitted to:

Name: Ahmed Imran Kabir

Designation: Lecturer, SoBE

Major: Management Information System

Submitted by:

Name: Abir Mahmud

Id: 111 202 156

Major: Management Information Systems

Registration Trimester: Spring 2024

School of Business and Economics

United International University

Date of Submission: July 20, 2025

Letter of Transmittal

Date: 20th July, 2025

Ahmed Imran Kabir

Lecturer, SoBE

United International University

Email: ahmedimran@bus.uiu.ac.bd

Subject: Submission of project report.

Dear Sir,

I am writing to properly submit my report on **Machine learning-based phishing detection from URLs**, which was assigned as a part of my BBA in Management Information Systems (MIS) requirement.

I believe the report accomplishes the goals stated in the project brief and offers understanding observations and recommendations. I hope you'll find it convenient and informative.

Thank you for the opportunity to work on this project, and please let me know if you have any questions or feedback.

Sincerely,

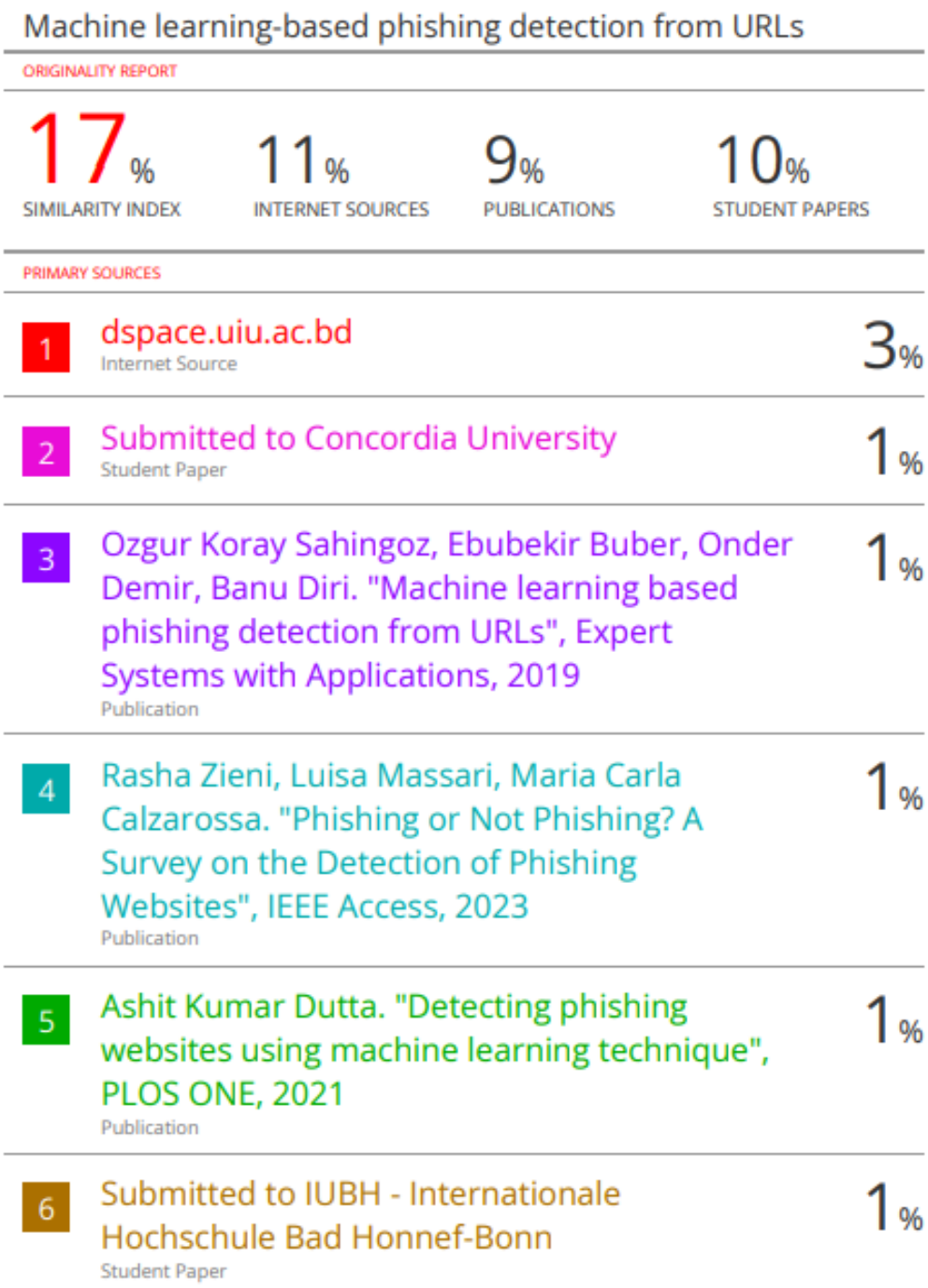
Abir Mahmud

ID: 111 202 156

BBA in Management Information Systems

United International University (UIU)

Certification Of Similarly Index



Machine learning-based phishing detection from URLs

ORIGINALITY REPORT

17%	11%	9%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.uiu.ac.bd Internet Source	3%
2	Submitted to Concordia University Student Paper	1%
3	Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, Banu Diri. "Machine learning based phishing detection from URLs", Expert Systems with Applications, 2019 Publication	1%
4	Rasha Zieni, Luisa Massari, Maria Carla Calzarossa. "Phishing or Not Phishing? A Survey on the Detection of Phishing Websites", IEEE Access, 2023 Publication	1%
5	Ashit Kumar Dutta. "Detecting phishing websites using machine learning technique", PLOS ONE, 2021 Publication	1%
6	Submitted to IUBH - Internationale Hochschule Bad Honnef-Bonn Student Paper	1%

7	arxiv.org Internet Source	<1%
8	Submitted to University of West London Student Paper	<1%
9	Ahmed A. Abd El-Latif, Mohammed A ElAffendi, Mohamed Ali AlShara, Yassine Maleh. "Cybersecurity, Cybercrimes, and Smart Emerging Technologies", CRC Press, 2025 Publication	<1%
10	Submitted to United International University Student Paper	<1%
11	Submitted to University of Cape Town Student Paper	<1%
12	Submitted to M S Ramaiah University of Applied Sciences Student Paper	<1%
13	arro.anglia.ac.uk Internet Source	<1%
14	Submitted to RMIT University Student Paper	<1%
15	Submitted to University of Hertfordshire Student Paper	<1%
16	Cheemaladinne Kondaiah, Alwyn Roshan Pais, Routhu Srinivasa Rao. "An ensemble learning	<1%

	approach for detecting phishing URLs in encrypted TLS traffic", Telecommunication Systems, 2024 Publication	
17	listens.online Internet Source	<1 %
18	pure.tue.nl Internet Source	<1 %
19	Submitted to Al-Qasim Green University Student Paper	<1 %
20	Submitted to University of Greenwich Student Paper	<1 %
21	Submitted to Liverpool John Moores University Student Paper	<1 %
22	Subrata Nath, Mohammad Manzurul Islam, Abdullahi Chowdhury, Mohammad Rifat Ahmmad Rashid et al. "Comprehensive Analysis of Feature Extraction Techniques and Runtime Performance Evaluation for Phishing Detection", 2023 6th International Conference on Applied Computational Intelligence in Information Systems (ACIIS), 2023 Publication	<1 %
23	Submitted to uwe Student Paper	<1 %

24	Submitted to University of Glamorgan Student Paper	<1%
25	ijirt.org Internet Source	<1%
26	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical and Computer Technologies", CRC Press, 2025 Publication	<1%
27	Submitted to New Era College Student Paper	<1%
28	jastt.org Internet Source	<1%
29	www.micromindercs.com Internet Source	<1%
30	www.ewubd.edu Internet Source	<1%
31	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dhirendra Kumar Shukla. "Artificial Intelligence, Blockchain, Computing and Security", CRC Press, 2023 Publication	<1%
32	Fatma Hendaoui, Saloua Hendaoui. "SENTINEY: Securing ENcrypted mulTI-party computatIoN for Enhanced data privacY and	<1%

	phishing detection", Expert Systems with Applications, 2024 Publication	
33	Submitted to Kingston University Student Paper	<1%
34	Submitted to University of Colorado, Denver Student Paper	<1%
35	aiforsocialgood.ca Internet Source	<1%
36	eprints.intimal.edu.my Internet Source	<1%
37	kuey.net Internet Source	<1%
38	vocal.media Internet Source	<1%
39	www.coursehero.com Internet Source	<1%
40	www.researchgate.net Internet Source	<1%
41	Zhang, Ruoyu. "Advancing Unobtrusive Wearable Technologies with Integrated Systems for Real-Time Physiological Health Monitoring.", University of California, Davis Publication	<1%
	iccs.ac.in	

42	Internet Source	<1%
43	telkomnika.uad.ac.id Internet Source	<1%
44	Submitted to Bath Spa University College Student Paper	<1%
45	Sally D. Abualgasim, Zeinab E. Ahmed. "chapter 2 Phishing Detection Methods", IGI Global, 2025 Publication	<1%
46	dokumen.pub Internet Source	<1%
47	etd.repository.ugm.ac.id Internet Source	<1%
48	web.archive.org Internet Source	<1%
49	www.hyperlearning.ai Internet Source	<1%
50	Gaurav Aggarwal, Ashutosh Tripathi, Himani Goyal Sharma, Tripti Sharma, Rishabh Dev Shukla. "Integrated Technologies in Electrical, Electronics and Biotechnology Engineering", CRC Press, 2025 Publication	<1%

51	Seok-Jun Buu, Sung-Bae Cho. "A Transformer Network Calibrated with Fuzzy Logic for Phishing URL Detection", Fuzzy Sets and Systems, 2025 Publication	<1%
52	free-soccer-picks.com Internet Source	<1%
53	repositories.nust.edu.pk Internet Source	<1%
54	Yassine Maleh, Mohammad Shojafar, Mamoun Alazab, Imed Romdhani. "Blockchain for Cybersecurity and Privacy - Architectures, Challenges, and Applications", CRC Press, 2020 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		

Declaration of the student

I hereby certify that, as a student, I have written this project paper honestly and to the best of my capability. This paper's content is all my own study and analysis, and all references have been appropriately declared.

I followed the instructions given by my supervisor, Ahmed Imran Kabir Sir, and finished the paper in the allotted time. I acknowledge that any sort of academic dishonesty is strictly forbidden and take full responsibility for the content in this paper.

Sincerely,

Abir Mahmud

ID: 111 202 156

BBA in Management Information Systems

United International University (UIU)

Acknowledgment

I want to sincerely thank my supervisor, Ahmed Imran Kabir Sir, for his invaluable advice, assistance, and encouragement during my research. His expertise and constructive feedback have been instrumental in shaping this project.

I also want to thank the BBA teachers for their helpful feedback and recommendations.

Finally, I am grateful to my family and friends for their unwavering love, support, and motivation.

Abstract

In these times, there has been a significant explosion of internet-connected devices, reaching from smartphones and IoT devices to cloud networks. Particularly, cybercriminals have increasingly turned their attention to these devices, employing phishing attacks that mark human weaknesses relatively exploiting organization weaknesses. So, a phishing attack, unsuspicious online users are cheated by apparently dependable entities into relating their personal data, such as login identifications and credit card details. This pilfered information becomes a valuable resource for scoring more sophisticated cyberattacks.

Although several researchers have future machine learning-based explanations to fight phishing attacks, these approaches often rely on a wide array of features, demanding large computational resources. This renders them impractical for devices with limited processing power. To tackle this challenge, the authors have developed a phishing detection method that successfully identifies phishing attacks using just nine lexical features. It showed tests using the ISCXURL-2016 dataset, covering 11,964 examples of genuine and phishing URLs. Their method was tested with various machine learning classifiers, achieving a remarkable accuracy rate of 99.57% when using the Random Forest algorithm.

[Keywords: Phishing, cybersecurity, Social engineering, Malicious URLs, Attack vectors, Deep Learning Techniques, Machine Learning Models, Feature Extraction, Word-Based Machine Learning, URL, Classification, Association Rule Mining, Convolutional Neural Network (CNN), Attention-Based, Hierarchical RNN, Automated Feature Learning, Phishing Threat Detection, Online Fraud, Sensitive Information Security, Cyber-Attack Tactics, Countermeasures, Multi-Factor Authentication, Sophisticated Phishing Schemes, URL Structure Analysis, Preventive Measures, Phishing Attack Channels, Natural Language Processing (NLP), Association Rule Mining, Random Forest Algorithm, Real-Time Phishing Prediction, URL Feature Analysis, Apriori Algorithm, Extrapolative Apriori, Transport Layer Security (TLS), Top-Level Domains (TLDs), Path Keywords.**]**

Table Of Contents

List of figures

List of Table

List of Acronyms & Abbreviations

URL - Uniform Resource Locators

MFA - Multi-Factor Authentication

NLP - Natural Language Processing

CNN - Convolutional Neural Network

RNN - Recurrent Neural Network

ML Model - Machine Learning model

SVM - Support Vector Machine

SQL – Structured Query Language

HTTP - HyperText Transfer Protocol

KNN - K-Nearest Neighbors

TLS - Transport Layer Security

SVM - Support Vector Machines

TLDs - Top-Level Domains

Chapter 1:

Introduction

1.1 Introduction

Phishing is a type of cyber-attack where attackers imitate legitimate entities to cheat individuals into providing sensitive personal information. Fraudulent emails that appear to be from reputable sources also fake websites designed to look like genuine ones. Nowadays, the main concern of security researchers is Phishing because it's not difficult to create a fake site that looks so close to genuine sites. Professionals can determine the fake websites, but some users may not be able to identify them and they become victims of phishing attacks. In order to get access to consumers' online accounts, hackers have installed malicious software in the computers and used systems that capture their user names and passwords. Phishers can good deal user data using a change of networks, such as email, Uniform Resource Locators (URL), instant chats, forum situations, phone calls, and text mails. Phishing content copies genuine content in structure and excites people to access it in order to get their searching data. Phishing is intended near fold some data for the purpose of financial gain or use of identity theft. Phishing is intended to obtain certain personal information for the purpose of obtaining money from or using identity theft. The global economy is being seriously damaged by malicious cyber-attacks. In addition, affording to the Challenging Phishing at Work Group's most new research arranged phishing patterns, financial payment organizations and webmail are a prime target of phishing attacks. The lack of users' awareness is a factor in the success of phishing attacks. Phishing attacks exploit the weaknesses that users have exposed, so it is very difficult to moderate them but they should be made even more active by enhancing detection techniques.

Criminals create illegal replicas of real Web sites and email accounts, usually from an economic organization or other organization production with finance information, to obtain private information. The logo and slogans of a real enterprise are used to create this email. Copying of images or a whole website may be enabled by the HTML design and structure. The fact that consumers are relying on the verification mechanism is one of the factors driving Internet's rapid growth as an information intermediate, and thus making it possible for them to use brands, symbols or more company identifiers. Phisher's sending "seized" messages to as several people as per it can in order to clasp users. When these

e-mails are opened, the customers tend to be unfocused from the legitimate entity to a spoofed website. Phishing in modern society is, for these reasons, extremely urgent, challenging, and critical. The assets of a domain, such as website URLs, website content, integrating both the website URLs and contented, the website source code, and the website screenshot, have been the subject quite a lot of recent studies against phishing. In order to protect users, however, there are no effective anti-phishing tools in an organization that could detect the malicious URL.

With the rapid advancements in global networking and communication technologies, many everyday activities such as social interactions, online banking, and e-commerce have increasingly moved to the Internet. This open, anonymous, and unrestricted nature of the Internet creates an ideal setting for cyberattacks, posing significant security threats not only to entire networks but also to individual users, including those who are technically skilled. Despite the vigilance and expertise of users, it is remarkably challenging to completely safeguard against falling victim to phishing scams (Greene, Steves, & Theofanos, 2018). Phishing attackers are becoming more sophisticated, taking into account users' psychological traits to deceive even the most experienced individuals (Curtis, Rajivan, Jones, & Gonzalez, 2018). These attackers craft their scams to exploit personal characteristics and behavioral patterns, making their schemes increasingly difficult to recognize and avoid. As a result, even knowledgeable users can be caught off guard by cleverly designed phishing attempts. The impact of such cyberattacks on end-users which means the consumer substantial. The consequences include significant financial losses and the waste of valuable time spent improving from these incidents.

Hackers are in control for billions of dollars in losses each year, highlighting the significant economic toll these cybercrimes impose on both individuals and businesses (Shaikh, Shabut, & Hossain, 2016). The financial impact of cyberattacks goes beyond immediate losses, surrounding expenses related to recovering from these incidents, such as repairing affected to the systems and addressing reputational harmless. The prevalence of phishing attacks and their efficiency in targeting users highlight the need for constant awareness and strong security measures. As attackers improve their strategies and exploit emotional management, it is essential for both individuals and organizations to

stay updated and accept urbane security practices. This includes educating users on phishing risks, improving detection methods, and using preventive strategies to lower the chances of submitting to these fake schemes.

The name of phishing stands up from the word fishing for victims, and detectives have paid close attention to this kind of attack. Attackers find phishing to be a gripping technique as it allows them to produce pretentious websites that look just like true ones. These websites vary from the standards in their Uniform Resource Locators (URLs), in spite of having like in writing user boundaries. By carefully investigative the URLs, an alert and informed user can usually detect these malicious sites.

However, due to the fast step of life, users often do not carefully check the URL of the page they are on, which may have been opened over other websites, social media, or emails (Gupta, Arachchilage, & Psannis, 2018). These deceiving URLs enable phishers to capture sensitive data such as financial data, personal details, usernames, and passwords. Users can unintentionally enter their data into these fake sites, believing them to be real.

Research on user identifications with phishing attacks specifies that users fall for phishing due to five main reasons:

- Lack of detailed information about URLs.
- Doubt about which web pages are reliable.
- Inability to see the full address of a web page due to redirection or hidden URLs.
- Lacking time to check the URL carefully, or inadvertent entry interested in a web page.
- Incapability to differentiate phishing websites from legitimate ones.

The Anti-Phishing Working Group reported on the prevalence of phishing attacks in the last quarter of 2016 (APWG, 2017), importance that such attacks mainly target users in developing countries, with China having the highest rate of infected computers at 47.09%, followed by Turkey at 42.88% and Taiwan at 38.98%. Moreover, the increasing use of

smartphones has led to less careful behavior while checking social networks on the go, making mobile users prime targets for attackers (Goel & Jain, 2018).

Several studies in the literature focus on detecting phishing attacks. Recent surveys discuss the general characteristics of phishing techniques, categorize the technical approaches used, and top some practical and effective opposing techniques (Chiew, Yong, & Tan, 2018; Qabajeh, Thabtah, & Chiclana, 2018).

Since phishing attacks exploit human weaknesses, additional support mechanisms are vital for defense. These mechanisms are classified into two main groups: increasing user awareness and employing additional software programs. Given the vulnerabilities of users, even skilled individuals can be deceived by new techniques, making software-based phishing detection systems preferable as decision support tools.

1.2 Increasing User Awareness:

1. **Educational Programs**: Regular training sessions and workshops to educate users about phishing tactics and how to recognize suspicious emails and websites.

2**. Simulated Phishing Attacks**: Conducting simulated phishing exercises to test and improve users' ability to detect phishing attempts.

3. **Awareness Campaigns**: Ongoing campaigns using emails, posters, and intranet messages to remind users of phishing dangers and best practices.

Using Additional Programs:

1. **Browser Extensions:** Developed or deploying browser extensions that can inform users of possibly dangerous sites or automatically block them.

2. **Email Filters:** Applying advanced email filtering systems that can detect and separation phishing emails before they spread the user.

3**. Machine Learning Systems**: Using machine learning models to analyze and classify URLs and web pages, refining the detection of phishing sites.

4. **Multi-Factor Authentication (MFA):** Requiring MFA for accessing sensitive information to add an extra layer of security, even if a user's credentials are compromised.

By merging user awareness creativities with strong software-based detection systems, organizations can create a multi-layered defense against phishing attacks, significantly reducing the risk and impact of these threats. A recent survey by Gupta et al. (2018) emphasized that new solutions to war phishing attacks often lead to the advance of new attack types exploiting these solutions' & weaknesses. Therefore, hybrid models are recommended completed single systems by network security administrators.

This paper focuses on real-time detection of phishing websites by investigating their URLs using different machine learning algorithms (seven implemented and compared) and various feature sets. The effectiveness of the proposed system depends not only on the dataset but also on the feature extraction process. We collected numerous legitimate and fraudulent URLs to construct our dataset, then defined three types of feature sets: Word Vectors, NLP-based, and Hybrid features.

The rest of the paper is organized as follows: the next section reviews related works on phishing detection. Section 3 discusses the challenges in detecting phishing attacks via URLs. Section 4 details the proposed system and dataset acquisition. Section 5 presents comparative experiments and results. The advantages of the proposed system are discussed in Section 6, followed by conclusions and future work in the final section.

1.3 Motivation

The most common method of social engineering attack is phishing. The phisher attempts to collect searching information from the user and complete such attacks to apply it fraudulently against the user or his or her company. Phishing schemes continue to become more sophisticated with targeted attacks posing a threat to many industries. Lance phishing is one of the more composite phishing systems that impends many trades. Many phishing emails are blocked by spam filters, but newer, trickier ones get through. There is proof that most people are aware that phishing attacks exist. In fact, a lot of businesses offer training and simulations to teach staff members how to know phishing

emails and messages. Despite this, scammers continue to utilize this type of cyberattack successfully, and it is still widely used. Additionally, fraudsters are adapting their strategies to defeat the anti-phishing security events in place.

To give an essence of how alarming the rate of phishing attack is, I have presented some information from a trusted source.

1	Every single day, over 15 billion spam emails are sent out, which resources that spam filters are "working overtime" and may allow damaging phishing attack emails to avoid them.
2	83% of trades said they were departed of phishing attacks in 2021. Six billion more doses are expected to occur in 2022.
3	About 214,345 distinct phishing websites were discovered in 2018, and since the beginning of 2020, the number of phishing attacks has increased by double.
4	30% of phishing emails are really unlocked. Accordingly, there is a advanced chance that someone will accidentally click on a damaging link or download a text that looks exciting but contains malware.
5	Approximately 90% of data breaches are the result of phishing. The US Federal Bureau of Investigation asserts that the annual growth rate of phishing attempts could reach 400%.

Table1: Phishing Facts and Statistics

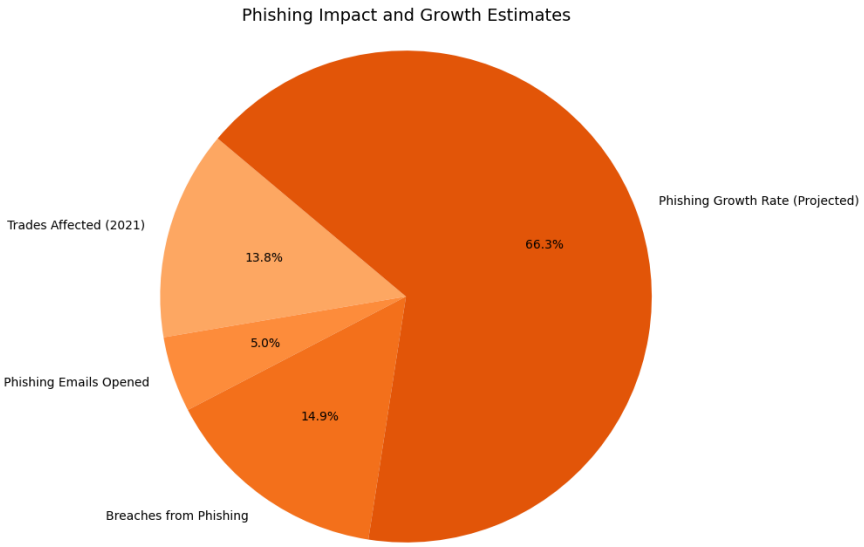


Fig 1: Phishing Impact Distribution

The graph depicting phishing sites from January 2007 to January 2021 reveals a concerning trend: the number of phishing sites has been growing exponentially. This sharp increase underscores the escalating threat posed by phishing attacks. To report this issue, here are three primary strategies:

1. Awareness Campaign: Educating users around phishing is vital in helping them identify and avoid potential pressures. Though, this method faces significant challenges. Largely spreading awareness and ensuring consistent user instruction across diverse populations can be difficult. Not all users may be attending or knowledgeable, which limits the efficiency of this strategy unaccompanied.

2. Blacklist: Applying blacklists involves keeping a list of identified phishing sites and blocking them to avoid user access. Even though effective to some amount, blacklists must have limitations. They require continuously updates to stay current with new phishing threats and may not cover all initial malicious sites. Additionally, this method relies on prior identification of threats on backlists which does not address new or unidentified phishing sites directly.

3. Machine Learning: Using machine learning offers a more flexible and scalable solution. Machine learning algorithms can evaluate large datasets to detect patterns and differences associated with phishing URLs. These models continuously learn from new data and adapt to growing phishing techniques. By examining various features of URLs, such as domain characteristics and URL structure, machine learning models can identify suspicious activity with greater accuracy and in real-time.

Given the limitations of toward know the awareness campaigns and blacklists, machine learning stands-out as a capable approach for phishing URL detection. It provides a dynamic and automated method to identify and mitigate phishing threats effectively. Our goal is to develop a machine learning model that can accurately detect phishing URLs, enhancing online security and offering users robust protection against fraudulent websites. This approach not only improves the ability to counteract current threats but also adapts to future phishing tactics.

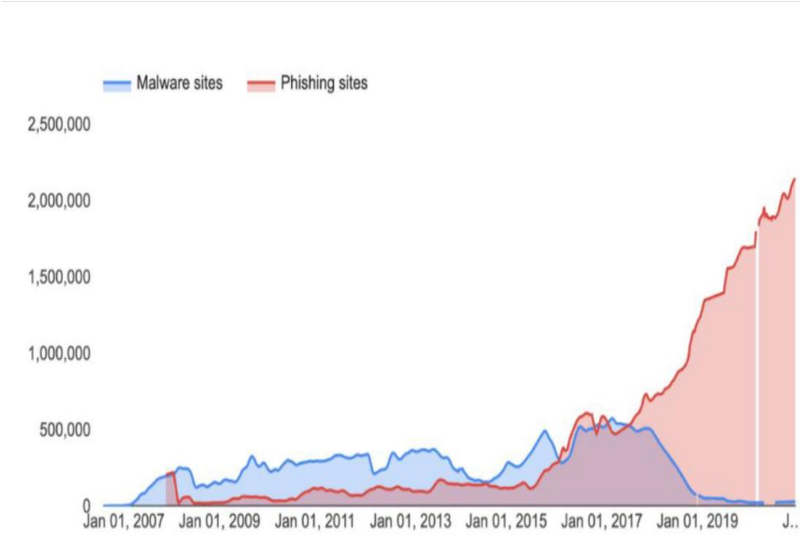


Fig 2: Trend of Malware and Phishing Sites Over Time.

1.4 Anatomy of phasing attacks:

As previously pointed out, phishing attacks are technically simple near implement. **(Figure-3)** drawings a representative situation of a phishing attack. In this situation, the attacker has two main roles, explicitly, making websites that expression very similar to the sites being copied and diffusion the consistent links using various communication channels, such as email or else social media. By connecting these links, individuals are directed to malicious websites where they strength end up figure-hugging sensitive data.

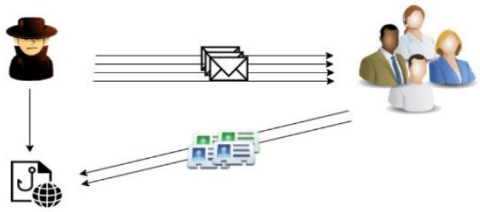


FIGURE 3. Main actors involved in a phishing attack.

Step-by-Step Breakdown of a Phishing Attack:

1. **Generating a Fake Website:**

Design and Layout: The attacker designs a website that strictly impressionists the target site, such as a bank, retailer, or social media platform. This includes copying the logo, layout, and overall appearance to make the site look authentic.

Domain Name Selection: Attackers often choose domain names that closely resemble the genuine site's URL. For example, they might use slight differences or add additional words that might not be noticed at a glance ("www.banksecure-login.com" instead of "www.bank.com").

2. **Disseminating the Malicious Links:**

Email Phishing: The attackers send out the mass email fake to be from a genuine organization. These emails often them that this cover urgent messages. Such as, account issues or security alerts, it motivates the addressee to click on a link.

Social Media Phishing: Attackers use social media platform to post or else send messages covering malicious links. These can look as if as promotions, messages from related persons friends, or even official posts from the organizations.

Other Channels: Phishing malicious links can also be spread over SMS (smishing), voice calls (vishing), immediate messaging apps, and even ads on legitimate sites.

3. Deceiving the Victim:

Landing Page: Once the victim connects on the phishing link then the phishers are sent to the fake website. The landing page generally contains a form asking for sensitive information, such as login identification, credit card details, also personal identification information.

Convincing Content: The fake website maybe included such as elements that add to its consistency, such as security seals, fake client reviews, and also other features typical of genuine sites.

Chapter 2:

Literature Review

2.1 Literature review

A regular literature review proceeding the categories based on phishing attack clarifications using deep learning methods.

This is the summary of current machine learning methods in order to detect phishing URLs that gives short information about how they have been applied. A total of 26 papers dealing with the same issue using various types of computer knowledge are presented in this area. In this way, we are looking at some of these 26 papers and learning how to spot phishing URLs.

A novel way of detecting phishing URLs through word-based machine learning in an immediate condition.

A method for grouping URLs according to some important parts such as protocol, domain, query, path and fragment has been proposed by the present paper. Following that, features like the number of @s and ?'s were calculated. They have predicted whether or not the URL is phishing based on these features, which use machine knowledge models. The precision obtained by this method can be seen in the following table.

Algorithms	Precision (%)	Confusion matrix	Recall (%)	F1 score (%)	Accuracy (%)
Random forest	99.7	[1937 11] [6 2039]	99.46	99.58	99.57
k-Nearest-Neighbor	98.67	[1937 11] [27 2018]	99.45	99.06	99.04
Support vector machine	96.87	[1918 30] [64 1981]	98.50	97.68	97.64
Logistic regression	94.96	[1874 74] [103 1942]	96.3	95.625	95.56

Table 2: Performance Comparison of Classification Algorithms on a Binary Classification Task

Intelligent phishing URL detection using association rule mining.

In this study, it established how to extract crucial elements from URLs, with feature work methods closely resembling those used in prior research. Though, this work applied association rule mining in place of the technique for predicting URLs. In difference, other studies such as those engaging CNN and attention-based graded RNN methods. Then focus on different approaches for detecting phishing URLs. Association rule mining

identifies patterns and relationships within URL data to predict likely phishing threats, while CNNs and attention-based RNNs influence deep learning methods to analyze URL features and rank their probability of being phishing attempts. By comparison these methodologies, the study highlights the effectiveness of association rule mining in identifying phishing URLs and also contribution an alternative to more complex deep learning models.

In this paper, introduces a model that participates CNN with an attention-based ranked RNN. Unlike methods that it involves physical feature mining, this model automatically learns how it is very essential features. The results show that this approach delivers excellent accurateness, outperforming numerous current high-tech machine learning models. The table below compares their results with individuals of top ML models, displaying the larger performance of their method.

Model	AUC	ACC	FPR	Precision
Random Forest on Handcrafted Features	0.92208	0.96892	0.00711	0.78360
Logistic Regression on Handcrafted Features	0.83802	0.96369	0.00082	0.94594
SVM on Handcrafted Features	0.83235	0.96781	0.00283	0.87793
Random Forest on Bag of Words	0.95515	0.94108	0.05829	0.45856
Logistic Regression on Bag of Words	0.94474	0.93292	0.06667	0.42655
SVM on Bag of Words	0.94726	0.97709	0.00430	0.90592
URLNet	0.94072	0.88447	0.12004	0.25544

Table 3: Model Performance Comparison

Phishing is a dominant form the online fraud in which attacker's skills fraudulent websites that closely mimic legitimate ones to deceive users into exposing sensitive information. These fake sites often replacement the arrival too functionality of famous, trusted websites, marking to trick those into entering personal details such as usernames, passwords, bank account information, or credit card numbers. Just because one of the older cyber-attack techniques, phishing remains highly effective and popular among cybercriminals due to its simplicity and broad reach. The important appeal of phishing lies in its manipulation of the essential trust users place in familiar-looking websites. Cybercriminals design these malicious sites to closely resemble legitimate ones, employing similar branding, logos, and design essentials to create a convincing front.

When users are focused to these fake sites, they may not proximately recognize the dishonesty, particularly if the phishing site look as if genuine and functions like the reliable site. As a result, users might unintentionally provide their searching information, intellectual they are networking with a trusted entity. The efficiency of phishing attacks is highlighted by the snowballing number of events reported annually. Numbers reveal a steady rise in phishing cases, reflecting the growing sophistication of attack methods and the persistent vulnerability of users to such schemes. The wide range of phishing attacks, joined with their ease of implementation, makes them a preferred choice for many attackers. Phishing can target individuals over various channels, including email, social media, and even text messages, further increasing its potential impact. The continuous flow in phishing incidents highlights the critical need for robust countermeasures. Organizations and individuals must remain attentive and employ comprehensive security practices to mitigate the risk of falling victim to phishing schemes. This includes educating users about recognizing phishing attempts, implementing advanced security technologies, and employing multi-factor authentication to protect sensitive information. As phishing tactics evolve and become more sophisticated, staying informed and proactive in defense tactics is essential to combating this ongoing threat effectively.

Even though progresses in cybersecurity, phishing persists as a significant threat because it takes advantage of on human error rather than technical liabilities. The ease of creating convincing fake websites and the wide reach of the internet contribute to the frequency and success of these attacks. As a result, phishing remains one of the most prevalent and effective tactics used by cybercriminals, highlighting the importance of ongoing awareness and preventive measures to protect sensitive information from being compromised.

Chapter 3:
Research Methodology

3.1 Methodology of the Study

3.1.1 System Architecture for Phishing URL Prediction:

A phishing URL is a web address created to deceive users into trusting they are directing to a genuine site but is actually designed to steal sensitive information, such as login IDs or financial details. These URLs often fake real websites by using similar domain names, a little different spellings, or visually marginal page designs.

Phishing URLs can be distributed through various channels, including email, social media, or nasty advertisements. They may prompt users to enter personal information on fake login pages or download malware. To identify and protect against phishing URLs, users should look for signs such as uncommon domain names, insecure HTTPS connections, and unexpected requests for sensitive information.

1. Data Collection and Preprocessing:

Data Collection: Phishing URL databases, which represent all safe and phishing URLs, are compiled.

Data preprocessing: Features are extracted from the URLs and the raw data is cleaned. This could entail analyzing URLs to extract parts such as the protocol, domain, query, path, and fragments as well as counting special characters or certain phrases.

2. Feature engineering:

From the obtained URL sections and other related information, important characteristics are generated using several tactics. This process aids in encoding URLs in a way that machine learning models can know them.

3. Development of Machine Learning Models:

Using the preprocessed dataset, several machine-learning models are created and trained. Chance Forestry, Pronouncement Tree, K-Nearest Neighbors (KNN), and other models are frequently employed. To improve the efficiency of the model, parameter adjustment can be done.

4. Evaluation Metrics:

The models' effectiveness in spotting phishing URLs is tested using important assessment indicators like precision, recall, precision, F1-score, and ROC-AUC.

5. Model Selection:

The top-performing machine learning strategy is picked and determined by the calculation outcomes. The random forest model got superior challenging precision in this case.

6. Deployment:

The model of choice is put into action for immediate time URL identification of phishing in an operating context. You may tool using a software interface or a web-based interface.

We deploy PyCham and Google Colab as our research and evaluation systems.

7. User Interface (online Application):

To enable users to input URLs for phishing detection, a simple online application has been developed.

Through a website interface, anyone can input URLs, and the model executed identifies that a given URL is phishing or not.

8. Results and Alerts:

The system shows users if a URL is safe or may indicate a fake website by providing outcomes.

A signal or notification is set up to alert the user or system administrator if a URL has been detected as phishing.

9. Feedback Loop:

The system may contain a user-feedback tool for analyzing input on how URLs should be identified. A model's precision can be further increased by using this feedback.

10. Monitoring and Maintenance:

The deployed system's performance and security are continuously observed upon. In order to keep updated on developing phishing strategies, frequently issued up-to-dates are carried out.

11. Future Improvements:

The results of deep learning may replace the requirement for individual characteristics in the future.

12. Impact on Society:

By assisting users and organizations in identifying and guarding against phishing assaults, the system lowers the danger of data breaches and monetary losses.

13. Security Measures:

To guard against assaults on the prediction model or the web application itself, the system may include security measures.

Therefore, this system building highlights every stage involved with applying machine learning to forecast phishing URLs, installing a basis for real-time identification, and maintaining protection. It requires for a powerful and profitable monitoring method in order to fight the growing risk of phishing attacks.

3.2 Complex Engineering Problems

P1	P2	P3	P4	P5	P6	P7
Complexity of Knowledge	**Range of Differing Requirements**	**Complexity of Analysis**	**Understanding of Issues**	**Degree of Applicable Codes**	**Degree of Stakeholder Involvement**	**Inter-need**
This project needs a study of current models with similar goals-(K8), data group and Pre-processing (K3, K6) & information for building a deep-learning model (K4,K5,K6)	For performance evaluation, we will use different evaluation metrics. There can be such scenarios where to Improve one metric we have to compromise another.	For data processing and extracting features, deep analysis is required.	✕	To build this project we have to follow the standards of making a website.	✕	✕

Table 4: Complex Engineering Problem Criteria

3.3 Complex Engineering Actives

A1	A2	A3	A4	A5
Range of Resources	**Level of Interaction**	**Innovation**	**Consequences for Society and Environment**	**Familiarity**
The project will require a wide range of resources, including people, different materials, information, and technology.	To resolve complex or conflicting technical, engineering, or other difficulties, students working on this project must collaborate effectively with each other.	✕	This project will have a favorable effect on society since it will help people to detect phishing URLs.	✕

Table 5: Engineering Activity Criteria

3.3.1 Implementation

In the implementation part I have established the model in the google colab. I had to preprocess the data first. And then extract the features out of it. The idea of feature extraction was taken from the papers that were reviewed.

3.3.2 Dataset

The dataset used here is the *'Phishstorm'* dataset. This is a balanced dataset covering almost 100000 examples. The dataset initially contains two columns URL and status. Where URL contains the URL and the status says it's genuine or phishing.

3.3.3 Models

The models that were used was Random Forest, Decision tree, KNN, LSTM+CNN and logistic regression. Out of all the models Random Forest showed the better accuracy.

3.3.4 Performance Evaluation Matrix:

Model	Train Accuracy	Test Accuracy
KNN	0.91	0.88
Pronouncement Tree	0.95	0.88
Casual Forest	0.95	0.90
Logistic Relapse	0.79	0.79
CNN + LSTM	0.76	0.71

Table 6: Model Accuracy Comparison

We can see, the best performance was shown by Random Forest. So, for deployment we have chosen the random forest.

The result shows the website where our model was deployed.

3.3.5 Future Extraction

Phishing attacks, one of the most dominant forms of cybercrime, remain to change, becoming more urbane and difficult to detect. The increasing complication of these attacks demands similarly advanced and adaptive defense mechanisms. Phishing attacks deceive individuals and organizations into revealing sensitive information such as passwords, financial details, or personal information, often through seemingly legitimate URLs and websites. As cheats constantly develop new strategies to evade detection, researchers and security experts are working to stay ahead by cultivating phishing detection methods. One promising direction in phishing detection research involves the design of machine learning algorithms capable of automatically extracting features from URLs. This approach can minimize or even remove the need for detailed, manual feature manufacturing, which traditionally has been a time-consuming and labor-intensive process. Rather than relying on human-defined rules to distinguish between legitimate and malicious URLs, these algorithms can learn and adapt from data. By doing so, they can detect subtle patterns and variations that may not be obvious to human analysts. The ability of machine learning models to autonomously gain insight into the characteristics of phishing URLs offers significant potential for improving detection accuracy. This approach allows predictive algorithms to evolve over time, adapting to new phishing techniques and URL variations as they emerge. By constantly learning from new data, machine learning models can identify previously unseen phishing strategies, making them more robust in their detection capabilities.

Also, to URL analysis, another exciting strategy involves the claim of natural language processing (NLP) to phishing detection. While traditional phishing detection methods often focus on URL structures or email headers, NLP can provide a profounder analysis of the content of phishing websites and emails. NLP models can examine the text on websites associated with suspicious URLs, detecting anomalies in language, tone, and structure that may indicate phishing attempts. For example, phishing websites often contain syntactic errors, unusual language usage, or content that differs subtly from the legitimate sites they are impersonating. By analyzing these linguistic clues, NLP-enhanced models could provide an additional layer of defense against phishing attacks.

The ongoing development of these advanced machine learning and NLP techniques must be attended by real-time data inputs and continuous model training. Phishing is a dynamic and rapidly evolving threat, with fraudsters continually devising new strategies to bypass existing security measures. Real-time data collection and model updates can help phishing detection systems remain up to date with the latest tactics employed by cybercriminals. By incorporating fresh data into the training process, models can quickly adapt to new phishing patterns and provide timely protection against emerging threats. A key challenge in phishing detection is the speed at which phishing attacks evolve. Fraudsters frequently change the URLs they use, create websites that appear increasingly legitimate, and use social engineering tactics to manipulate users into falling for scams. As these techniques become more sophisticated, detection tools must also become more agile and adaptive. In this regard, the future of phishing detection will likely involve a combination of innovative technologies and novel strategies that enhance the ability of defense systems to keep pace with cybercriminals.

One such strategy may involve the use of deep learning models that are capable of processing vast amounts of data and identifying complex, non-linear patterns in phishing attacks. These models could participate multiple sources of information, including URL structure, website content, email metadata, and user behavior, to provide a more comprehensive assessment of potential phishing threats. By considering a broader range of factors, deep learning models can offer a more nuanced and accurate detection system that is difficult for fraudsters to evade. Also, as phishing attacks become more combined with other types of cybercrime, such as malware distribution and ransomware attacks, future phishing detection systems will need to incorporate broader cybersecurity frameworks. By joining phishing detection with other aspects of cybersecurity, such as malware detection or threat intelligence, organizations can create a more holistic approach to defense. This integrated system could provide real-time alerts and coordinated responses to multiple types of cyber threats, ensuring a more resilient defense against increasingly complex attacks. Ultimately, the next phase of phishing detection will depend on the continuous advancement of technology and the development of new methodologies. Machine learning, natural language processing, deep learning, and real-time data inputs are just a few of the tools that will shape the future of phishing

detection. As phishing attacks continue to grow in complexity, staying one step ahead of cybercriminals will require not only innovative technological solutions but also a proactive and adaptive approach to security.

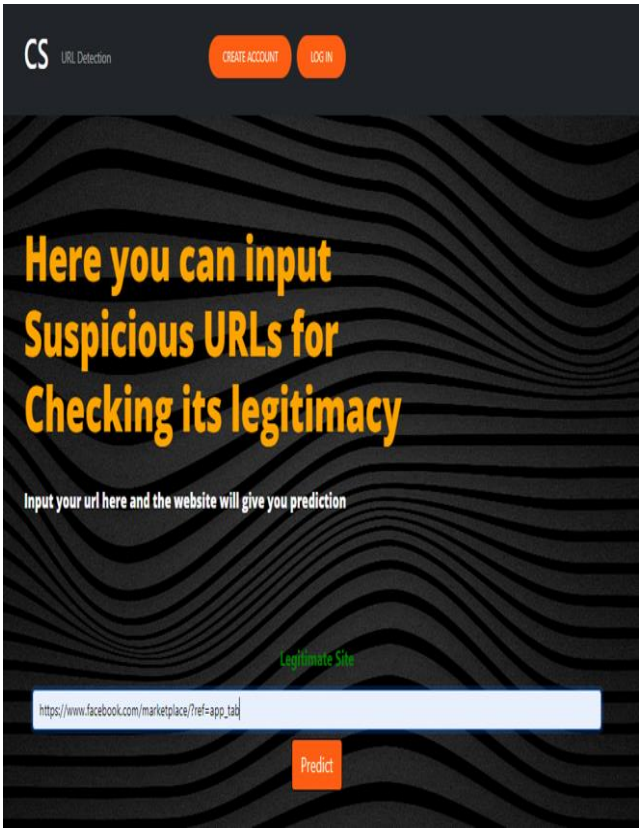


Fig 4: URL Legitimacy Checker

3.3.6 Association rule mining to detect phishing URL

Association rule mining can be effectively utilized to detect phishing URLs by analyzing significant patterns and relationships within URL data. This approach focuses on identifying frequent associations and links between various components of URLs, such as domains, paths, and parameters. By examining these patterns, researchers can uncover specific character sequences or combinations that may indicate phishing attempts. In practice, association rule mining involves breaking down URLs into their constituent parts and finding frequent item sets of URL components that frequently co-occur. For example, certain patterns involving unusual symbols, specific character sequences, or multiple subdomains are often associated with phishing attempts. By identifying these patterns, association rule mining helps in flagging URLs that deviate

from legitimate structures, thereby identifying potential phishing sites. Although association rule mining was not the primary focus of our project, its application in previous studies has demonstrated its utility in phishing detection. The method provides valuable insights into common features and structures of phishing URLs, complementing other machine-learning techniques. While our project primarily employed different machine learning methods, incorporating association rule mining can enhance phishing URL detection by offering an additional layer of analysis and revealing underlying patterns in URL data.

Chapter 4:
Experimental Results

4.1 Review of Prior Research

In this section, we explore various important studies and existing approaches in the field of phishing attack detection as documented in academic literature. Phishing attack detection methods can be categorized into two main groups: those based on educating users and those reliant on software-based solutions. The software-based approaches can be further divided into four categories: blacklisting, visual similarity analysis, machine learning, and hybrid methods. We will offer a comprehensive overview of these existing works shortly.

Phishing attacks remain a significant threat in the realm of cybersecurity, necessitating robust detection mechanisms.

User Education-Based Methods:

User education efforts on training individuals to identify and avoid phishing attempts. This approach is created on the idea that informed users are less possible to fall victim to phishing tricks. Educational creativities can include workshops, sessions, online courses, also cooperative training programs. These programs often cover topics such as identifying suspicious emails, identifying deceptive URLs, and understanding the tactics normally used by phishers.

Despite their standing, user education methods have restrictions. The efficiency of these programs depends mainly on user participation and retaining of the information. Also, as a phishing system grow, constant instruction is required to keep users updated on new threats. Though, user education remains a serious component of a full phishing defense strategy.

Software-Based Solutions:

Software-based solutions are force technology to detect and block phishing attacks. These approaches can be more divided into four categories: blacklisting, visual similarity analysis, machine learning, and hybrid methods.

1. Blacklisting:

Blacklisting includes maintaining and updating a list of identified phishing sites. When users attempt to visit these sites, they are blocked or warned. This method is straightforward and easy to device, making it a popular choice. Though, its efficiency is limited by the time lag between the identification of new phishing sites and their addition to the blacklist. Moreover, new phishers can avoid blacklists by frequently changing their URLs.

2. Visual Similarity Analysis:

Visual similarity analysis detects phishing sites by comparing them to genuine websites based on visual elements such as design, logos, and color systems. This method accepts that phishing sites often mimic the appearance of legitimate sites to deceive users. Techniques such as image processing and feature mining are used to identify visual similarities. While this method can be effective, it requires significant computational resources and may produce false positives if legitimate sites experience design changes.

3. Machine Learning:

Machine learning uses procedures to identify phishing patterns and forecast potential attacks. By training on large datasets of legitimate and phishing websites, machine learning models can acquire to differentiate between the two. Features used in these models can include URL characteristics, HTML content, and website behavior. Machine learning approaches offer high truth and flexibility, as models can be updated with new data to recognize developing phishing techniques. However, the quality of the detection rests on the quality and diversity of the training data.

4. Hybrid Methods:

Hybrid methods combine several techniques to improve detection accuracy and strength. For example, a hybrid approach might participate blacklisting with machine learning to force the strengths of both methods. By joining different detection mechanisms, hybrid methods can achieve higher detection rates and reduce false positives. These approaches are mainly useful in addressing the developing nature of phishing attacks.

The use of phishing detection methods varies based on their implementation and the context in which they are used. User education, while crucial, needs to be complemented by robust software-based solutions to provide full protection. Blacklisting offers simplicity and ease of use but struggles with adaptability. Visual comparison analysis provides a more nuanced detection device but requires large computational power. Machine learning offers adaptability and high accuracy but depends on good training data. Hybrid methods represent a balanced approach, leveraging the strong point of several techniques to enhance detection capabilities.

Future research in phishing attack detection should focus on improving the adaptability and truth of existing methods, reducing false positives, and developing integrated solutions that combine user education with advanced software-based techniques. By addressing these areas, researchers and experts can better protect users from the ever-evolving threat of phishing attacks.

4.2 Objectives of the Design

The primary aim of this research is to grow a machine learning-based system for efficiently detecting phishing challenges. The fundamental goal is to enable users to fast measure the legitimacy or wickedness of a given URL with minimal time and effort. To achieve this, our approach involves several key objects.

1. Feature Extraction and Pre-processing:

The first step in our approach is the extraction of feature vectors from a URL when a user entree it. Feature extraction is crucial because it involves identifying and measuring various mechanisms of the URL that may indicate phishing. These mechanisms can include the length of the URL, the occurrence of certain keywords, the structure of the domain, the use of different types, and the frequency of numerical values.

When these feature vectors are extracted, they are pre-processed to ensure they are in a appropriate format for analysis by machine learning algorithms. Pre-processing steps can include normalization, which adjusts the feature values to a common scale, and encoding categorical variables into numerical format. This step is vital for attractive the accuracy and act of the machine learning models.

2. Machine Learning Algorithm Execution:

In the process of phishing URL detection, machine learning algorithms play a crucial role in categorizing URLs based on their legitimacy. After preprocessing and feature extraction, the next step involves feeding the prepared feature paths into various machine learning algorithms. The choice of algorithms is critical, as different methods offer separate advantages and contribute to the system's overall effectiveness.

Decision Trees are a popular choice for their interpretability and simplicity. They build a model that swift data based on feature values, making it easy to understand how decisions are made. However, they can be likely to appropriate, particularly with complex datasets. **Random Forests**, an ensemble method that combines multiple decision trees, address the overfitting issue by averaging the predictions of individual trees. This approach improves accuracy and simplifies better to new data, making it a strong option for phishing detection.

Support Vector Machines (SVM), are effective in solid spaces and are known for their strong performance in classification tasks. SVMs work by finding the optimal hyperplane that separates different classes, which can be particularly useful for distinguishing between phishing and legitimate URLs. Neural Networks, particularly deep learning models, excel at taking complex patterns in large datasets. They consist of multiple layers that learn classified features, allowing them to detect subtle hints in URL characteristics. Despite their computational intensity, neural networks can significantly improve detection accuracy.

Each algorithm brings unique strengths to the table, and the choice of which to use may depend on the specific characteristics of the dataset and the desired balance between interpretability and performance. Combining or comparing these algorithms through research can lead to a more robust phishing detection system.

3. Real-Time Operation:

The main goal of the proposed system is to ensure real-time functionality, allowing users to receive immediate feedback on the legitimacy of URLs. This real-time capability is critical; as phishing attacks often depend on users reacting quickly to deceptive content. By instantly evaluating and flagging potentially malicious URLs, the system can greatly reduce the likelihood of successful phishing attempts. To achieve this, the system must be carefully optimized for rapid data processing with minimal delays. This involves implementing efficient data structures that streamline operations and minimize processing overhead. Additionally, parallel processing will be utilized to manage multiple tasks simultaneously, speeding up the analysis of large volumes of URL data. Advanced computational methods, such as machine learning algorithms and heuristic analysis, will also be crucial in enabling the system to quickly and accurately determine the legitimacy of URLs. By combining these strategies, the system aims to offer strong protection against phishing threats, ensuring that users receive timely warnings and can make informed decisions to protect their online interactions. Ultimately, the system's real-time capabilities are key to its success in addressing the immediate and ever-changing nature of phishing attacks.

4. Optimized Algorithms and Efficient Data Processing:

To ensure minimal latency and provide instant results, the system employs a combination of optimized algorithms and efficient data processing techniques. The optimization process involves fine-tuning algorithms by adjusting parameters to improve performance, making them more responsive and faster. This fine-tuning is essential for reducing the computational burden and ensuring that the system can operate swiftly even when handling large volumes of data. Additionally, the system makes use of lightweight models designed to deliver high accuracy while requiring less computational power, which further enhances speed and efficiency.

Efficient data processing techniques are also a key component of the system's design. For instance, batch processing is used to manage multiple URLs simultaneously, allowing the system to quickly analyze large datasets in parallel rather than sequentially. This approach significantly reduces the time needed to process each URL. Furthermore, the

system leverages in-memory databases for rapid data access and retrieval. In-memory databases store data directly in the system's RAM, enabling faster data processing compared to traditional disk-based databases. This combination of algorithm optimization and advanced data processing techniques ensures that the system can deliver real-time results, providing users with timely and accurate feedback on the legitimacy of URLs.

5. Comprehensive Detection Mechanism:

In addition, to real-time detection, the system is designed to offer a thorough approach to identifying phishing attempts by integrating various detection methods. This comprehensive strategy includes heuristic analysis, blacklist checking, and behavioral analysis, each contributing to the overall effectiveness of the system.

Heuristic analysis focuses on examining the specific characteristics and patterns within a URL that are typically linked to phishing activities. This method helps to identify suspicious URLs based on known traits of phishing sites. Meanwhile, blacklist checking involves comparing the URL against a database of known malicious sites, immediately flagging any matches as potentially dangerous. Behavioral analysis adds another layer of protection by monitoring the actions taken by the URL after it is accessed. This includes tracking activities like redirections or the execution of scripts, which are common indicators of phishing behavior. By analyzing these behaviors, the system can detect and respond to suspicious activity in real time. These integrated methods create a robust mechanism for identifying phishing attempts, ensuring users are protected from a wide range of threats.

6. User-Friendly Interface:

Another crucial objective is to develop an accessible interface that simplifies the calculation of URL safety for users. This interface must be designed to deliver clear, actionable feedback, allowing users to quickly determine if a URL is safe or potentially dangerous. Including visual elements such as color-coded alerts and risk scores increases user understanding by providing direct, innate signals about the URL’s safety. For example, green can denote safe URLs, while red or yellow might indicate probable threats, helping users make up-to-date decisions at a glance. Visual indicators, the

interface should offer detailed explanations for URLs identified as suspicious. This educational component is essential for helping users understand the specific reasons ahead of the assessment. Explanations could include information on unusual URL patterns, suspicious components, or how certain features contribute to the supposed risk. By providing this context, users gain insight into the factors that led to the detection of phishing risks. Such educational feedback not only aids in immediate decision-making but also empowers users with the knowledge to better identify potential threats in the future. For example, understanding common phishing signs helps users become more aware and discerning when encountering new URLs. Overall, the interface should strike a balance between simplicity and depth, ensuring that users can easily navigate the calculation process while also acquisition valuable insights into URL safety. This approach improves user experience and enhances the overall effectiveness of the phishing detection system.

So, the interface should strike a balance between simplicity and informative depth, offering both quick visual cues and in-depth explanations. This approach not only ensures that users can easily navigate the assessment results but also empowers them with the knowledge to protect themselves from phishing attempts in the long run.

7. Scalability and Adaptability:

The system is concocted to be both scalable and adaptable, enabling it to effectively manage the ever-evolving landscape of phishing attacks. Scalability is a critical feature, ensuring that the system can handle an increasing volume of URLs without sacrificing performance or speed. As the number of URLs to be processed grows, the system must maintain its efficiency, allowing it to deliver real-time results even under heavy loads. Adaptability, on the other hand, is essential for keeping the system up-to-date with the latest phishing tactics. Phishing techniques are constantly changing, and the system must evolve alongside them. This involves regularly updating the machine learning models and detection rules to identify new and emerging threats. A key aspect of this adaptability is the incorporation of a continuous feedback loop. As the system encounters new data, it learns from these experiences, refining its algorithms and improving its detection capabilities over time.

By combining scalability with flexibility, the system ensures long-term effectiveness in identifying phishing attempts. It can grow to meet increasing demands while staying agile enough to respond to new threats, ultimately providing robust and reliable protection for users against a wide range of phishing attacks.

In summary, the objectives of the design are centered on developing a machine learning-based system that efficiently detects phishing attempts in real-time. By focusing on feature extraction, machine learning implementation, real-time operation, optimization, comprehensive detection, user-friendly interface, and scalability, the system aims to provide users with a reliable tool for assessing the legitimacy of URLs. This holistic approach ensures that users are protected from phishing attacks with minimal effort and maximum accuracy.

Chapter 5:

Discussion

5.1 Conclusion

This paper focuses on analyzing URL features through the application of associative rule mining techniques, particularly using the apriori and extrapolative apriori algorithms. The rules generated from these techniques provide valuable insights into the common characteristics of phishing URLs. The study identifies several key indicators that are crucial for detecting phishing URLs. Among these indicators are the existence of Transport Layer Security (TLS), which can sometimes be misleading in phishing attempts, and the appearance of specific top-level domains within the URL, which can be a red flag.

Additionally, the study emphasizes the significance of certain keywords embedded within the URL's path, which are often related to phishing sites. It also highlights other critical factors that contribute to the detection of phishing URLs, such as the total number of slashes in the URL, the presence of dots in the host section, and the overall length of the URL. These elements are particularly telling when distinguishing between legitimate and malicious URLs. By analyzing these features, the study offers a comprehensive approach to identifying phishing attempts and enhancing the accuracy and effectiveness of phishing detection methods.

Reference Link APA:

1. Zouina, M., & Outtaj, B. (2017). A novel lightweight URL phishing detection system using SVM and similarity index. Human-centric Computing and Information Sciences, 7, 1-13.

2. Aung,%20E.%20S.,%20Zan,%20C.%20T.,%20&%20Yamana,%20H.%20(2019).%20A%20survey%20of%20URL-based%20phishing%20detection.%20In%20DEIM%20forum%20(pp.%20G2-3).

3. Aljofey, A., Jiang, Q., Qu, Q., Huang, M., & Niyigena, J. P. (2020). An effective phishing detection model based on character level convolutional neural network from URL. Electronics, 9(9), 1514.

4. Asiri, S., Xiao, Y., Alzahrani, S., Li, S., & Li, T. (2023). A survey of intelligent detection designs of HTML URL phishing attacks. IEEE Access, 11, 6421-6443.

5. Huang, Y., Yang, Q., Qin, J., & Wen, W. (2019). Phishing URL Detection via CNN and Attention-Based Hierarchical RNN. 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), 112-119.

6. Yang, R., Zheng, K., Wu, B., Wu, C., & Wang, X. (2021). Phishing website detection based on deep convolutional neural network and random forest ensemble learning. Sensors, 21(24), 8281.

7. Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, *117*, 345-357.

8. Singh, S., Singh, M. P., & Pandey, R. (2020, October). Phishing detection from URLs using deep learning approach. In 2020 5th international conference on computing, communication and security (ICCCS) (pp. 1-4). IEEE.
9. Das Guptta, S., Shahriar, K. T., Alqahtani, H., Alsalman, D., & Sarker, I. H. (2024). Modeling hybrid feature-based phishing websites detection using machine learning techniques. Annals of Data Science, 11(1), 217-242.
10. Korkmaz, M., Sahingoz, O. K., & Diri, B. (2020, July). Detection of phishing websites by using machine learning-based URL analysis. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.
11. Yang, P., Zhao, G., & Zeng, P. (2019). Phishing website detection based on multidimensional features driven by deep learning. IEEE access, 7, 15196-15209.