ISSN 2047-3338

Mitigating Malicious URLs Using Machine Learning Techniques

MeharunNisa[1], Ahmad Raza[2], M. Junaid Arshad[3]

[1-3]Department of Computer Science, University of Engineering and Technology, Lahore, Pakistan

[1]meharunisa235@gmail.com, [2]m.ahmadraza457@gmail.com

***Abstract*—This paper provides a review of recent studies on malign URL detection using models from Machine Learning, with a focus on their application in the context of the what dataset being used, what number of dataset being included, if feature engineering has been applied and if yes then what type of features are being used, also the comprehensive view of classifiers used at time in all those papers under discussion with. Moreover, it has analyzed each study's methodology, outcomes, and conclusions, and discuss the strengths and limitations with mentioned future work in each. While the tabular comparisons are for the very latest publications as of 2023 and 2024. The analysis shows that while existing metrics can be useful for identifying areas of improvement, they may not always provide a complete picture of software quality. The goal of this study is to determine the finest factors helping in getting the effective technique for spotting a phishing URL in large datasets. When employing machine learning algorithms to identify phishing URLs, users encounter numerous difficulties. It must be done to defend users against phishing attempts if you want people to continue having trust with online platforms and services. In order to ensure the security of user details as well as adhere with industry standards and data protection requirements, phishing URL detection must be reliable.**

***Index Terms*—Machine Learning, Intrusion Detection, Cyber Attack and Phishing Detection**

I. Introduction

IN today's world, data security is a top priority for all of us when our personal data is being accessed by almost all the websites or apps we use in our mobiles, laptops etc. Irrespective of all security steps, theft of personal credentials can be done using various methods and very common from them are like pop ups on websites redirecting to malicious web pages, instant messaging on social media platforms like people send URLs of gifts to click and users click them to get some benefit, but they lead to security leakages. Malicious URLs can also be distributed through mobile devices, such as smartphones or tablets, through phishing messages, malicious apps, or mobile malware. Phishing emails often contain links to malicious websites or web pages that mimic legitimate websites to steal users' credentials or personal information. Malicious domain names may be misspelled versions of legitimate domain names or may contain words or phrases that are designed to trick users into clicking on them.

Malicious URLs and domain names are a common way for attackers to distribute malware, steal personal information, or conduct phishing attacks. The steps have been taken by individuals and organizations by improving security parameters such as the firewalls, also the antivirus of software, and the unauthorized action detection systems, to prevent cyber-attacks and mitigate the impact of any successful attacks. Education and awareness campaigns have been launched to help users identify and avoid malicious URLs and domain names. Users are encouraged to be cautious when clicking on links or opening email attachments, and to get surety that their in-use devices and software are always updated with the latest security measures. Governments and regulatory bodies have introduced laws and regulations to improve cybersecurity and protect against cyber-attacks. Internet service providers (ISPs) and domain name registrars have implemented improved infrastructure to help prevent the registration of malicious domain names and to track down and disable any that are discovered.

Whereas Machine learning algorithms have been trained to analyze and identify patterns in large datasets of known malicious URLs. These patterns could include things like the use of certain words or characters, or the presence of certain types of domains or subdomains.

Once the machine learning algorithm has been trained on these datasets, it can then be used to scan new URLs and determine whether or not they are likely to be malicious. This is done by comparing the characteristics of the new URL to the patterns that the algorithm has learned to associate with malicious URLs.

The more data that the algorithm is trained on, the more accurate it becomes at detecting malicious URLs. Additionally, machine learning algorithms can be updated and refined over time as new threats emerge and the characteristics of malicious URLs evolve.

The remainder of the paper starts with a presentation of literature review (Section II). It is followed by methodology

(Section III). Finally, a conclusion is drawn (Section III).

II. Literature Review

The literature review is presented in tabular form, comparing various papers across different columns. It provides a concise summary of the key findings and highlights the similarities and differences between the papers. This format allows for easy comparison and analysis of the literature. There is a comparison of datasets number used in reviewed papers and also the source of datasets, this will be giving a quick view of how much data has been used for evaluation of classifiers in proposed solutions by the papers.

In paper [1] attributes extracted on the basis of lexical and host based and content-based features from each URL, then passed to classifier to check if its clean or not. They extracted 54 attributes. The findings of the referenced article have been utilized to develop a freely available tool for identifying malicious URLs on web browsers. *Gap:* To address the identified gap, readers are motivated to explore other algorithms such as Naïve_Bayes, Decision_trees, k_nearest_neighbors, neural_networks, etc., in addition to the proposed method. *Future work:* In terms of future work, the outcomes of this research hold potential for practical implementation in infor- mation security systems, enhancing their overall effectiveness. In [2] 62 features have been selected from categories as Domain based features, Host_based, Reputation_based, and Lexical features. *Gap:* The mentioned gap indicates that K-SVM and CNN algorithms were unable to find solutions within a time frame of 7 days, while the Voted Perceptron algorithm faced memory limitations when processing the RoBERT dataset. *Future work:* Looking ahead, future research will focus on improving the classifiers in terms of time and accuracy calculated in training, leveraging the unique characteristics of URLs which are encoded in graphical quick response (QR) codes, and enhancing the detection rate for quick response (QR) code frauds.

The paper in comparison authored by DR. S. K Singh, Poonam Jain, Ritesh Mourya, Ahmad Khan [4] used Extra Tree classifier that differentiates it from others which concluded the accuracy result for the mentioned classifier as 80.67%. They used maximum dataset instances and also cleaned it in preprocessing to remove null values and duplicates which reduces the dataset from 650000 to 620000 urls. Under sampling and Over sampling has been applied. *Future work:* Authors mentioned in the paper involves enhancing the detection of phishing attempts by breaking down URL structures, examining user behavior, looking for anomalies in web or email content, researching evasion strategies, putting adaptive algorithms in place, and enhancing model transparency for user confidence.

Habiba Bouijij and Amine Berqia in [5] mainly focused on dataset and used two different distinct datasets of URLs. One is collected from 2016 and the second is of 2021. Their types of feature extraction variety were larger in number and hence shown 99% accuracy with dataset from 2016 while 90.37% for the other one. *Future work:* Authors aim to advance phishing the identification through development of a model capable of navigating difficult URLs and picking up on minute details. So is thought by adding additional techniques—such as URL HTML Encoding, WHOIS method, Tiny URL technique, and an advanced voting technique—will boost detection strength and validity.

Another paper from 2024, in which R. Jayaraj, A. Pushpalatha, K. Sangeetha, T. Kamaleshwar, S. Udhaya Shree, Deepa Damodaran [6] proposed a method using Hybrid Ensemble Feature Selection that combines ensemble learning and feature selection. It chooses the most appropriate characteristics for prediction based on the variety of several models in an ensemble. By keeping only the most useful characteristics from the original dataset, this method raises clarity, decreases over fitting, which improves model performance. The result revealed an accuracy of 97.6%.

Usha Ruby in their paper [7] used 12000 malicious and 39000 normal dataset and used features in feature extraction in large number with recursive elimination for feature selection with k means classifier shown high results out of others. Recursive elimination works by recursively removing features and building a model on the remaining features until the desired number of features is reached or the model performance no longer improves. The classifier used is FSRE-K-means-CNN stands for Feature Selection and Reduction Ensemble with K-means Clustering and Convolutional Neural Network. It's a method that combines feature selection, feature reduction, and ensemble learning techniques, along with K-means clustering and convolutional neural networks (CNNs), to improve the performance of classification tasks. And performed performance Analysis of FSRE-K-means-CNN across various URL Lengths like 10 to 20, 40 to 50 and so on to 200. In result high accuracy being observed as 99%. While didn't mention any type of feature extraction and future work.

Ayan Mahmood, Vishal Pandey, Rohit Raj, Gouri Shankar Mishra [8], used 5 models and splitted the dataset in 5 sets. They trained separate dataset for each model that can increase the accuracy. [8] explained the importance and working of each model very well for the readers to understand and also brought the results where XGBoost had the highest accuracy. *Future work:* includes to examine how deep learning can be integrated to extract features from text and webpage code more effectively. They also intend to develop a browser plugin that would allow them to easily integrate their approach into web browsers.

In paper [9] the method detects malicious traffic in proxy logs by analyzing 10-line paragraphs containing malicious URLs using a supervised learning model. *Gap:* This method has not yet been applied to actual proxy server logs. *Future work:* applying the method to actual proxy server logs is a future work. One improvement plan is adjusting the size of a paragraph. Another plan is using other NLP techniques to summarize a paragraph. In paper [9] authors have used methodology that is based on the analysis based on breaking characters as lexical analysis and later assigning numerical values to them which are referred as feature

Fig. 1: Dataset Number and Sources

Title	Year	Paper Type	Dataset Type							Dataset	
			Domain names	Proxy Server Logs	Pcap Files	Login URLs	Web App URLs	IP address	resource path	total	
										number	source
Phishing Website URL's Detection Using NLP and Machine Learning Techniques	2023	Journal on Artificial Intelligents	✗	✗	✗	✗	✓	✗	✗	550000	N/A
Classification of Malicious URLs Using Machine Learning	2023	Sensors - scientific journal - by MDPI	✗	✗	✗	✗	✓	✗	✗	650000	Kaggle Repo
Detecting Phishing Domains Using Machine Learning	2023	Applied sciences - by MDPI	✗	✗	✗	✗	✓	✗	✗	11055	University of California, Irvine UCI Phishing Websites
Phishing URL Detection Using Machine Learning	2024	IRE Journals	✓	✗	✗	✗	✓	✓	✗	620000	kaggle, phishtank, Aalto University UrlSet
Phishing Website Classification using Machine Learning with Different Datasets	2024	International Journal of Computing and Digital Systems	✓	✗	✗	✗	✓	✗	✗	20000	2016 - Canadian Institute for Cybersecurity, 2021 - Mendeley Data
Intrusion detection based on phishing detection with machine learning	2024	Journal of Measurement-sensors	✗	✗	✗	✗	✓	✗	✗	3000	N/A
Enhancing Phishing URL Detection Accuracy in Software-Dened Networks (SDNs) through Feature Selection and Machine Learning Techniques	2024	Research Square	✓	✗	✗	✗	✓	✗	✗	51100	phishtank, 5000best
Detection of Phishing Sites Using Machine Learning Techniques	2024	International Research Journal on Advanced Engineering Hub	✓	✗	✗	✗	✗	✓	✗	Not Mentioned	N/A

The Tabular comparison is for papers in 2023 and 2024 which are [1] – [8].

Fig. 2: Tools and Libraries used

Title	Year	Tools/Libraries
Phishing Website URL's Detection Using NLP and Machine Learning Techniques	2023	Python 3.8, sci-kit-learn 0.24, TensorFlow 2.4, NLTK 3.5 Development tool—Data lore JetBrains
Classification of Malicious URLs Using Machine Learning	2023	Matlab 2022
Detecting Phishing Domains Using Machine Learning	2023	Not Mentioned
Phishing URL Detection Using Machine Learning	2024	Not Mentioned
Phishing Website Classification using Machine Learning with Different Datasets	2024	Not Mentioned
Intrusion detection based on phishing detection with machine learning	2024	Python Weka
Enhancing Phishing URL Detection Accuracy in Software-Dened Networks (SDNs) through Feature Selection and Machine Learning Techniques	2024	Open Network Operating System controller, Mininet tool
Detection of Phishing Sites Using Machine Learning Techniques	2024	Not Mentioned

quantification for the detection or to catch the malign domain names and then those names will be comparing with other two real life malign domain names detection models.

Gap: This paper proposed that using features as lexical and characters distribution does not seem comprehensive and for that reason it cannot detect all the malicious or malignant domain names present on the Internet, but in case if the malicious or malign domain names are generated randomly, then this approach can detect them more efficiently. *Future work:* They will be working for the further improvement of the methods presented, and a more in-depth analysis with the use of data sets in extensive or large size. Numerous studies have been conducted to identify and prevent SQL injection attacks. In this section, we will provide a comprehensive review of recently published research papers that are relevant to this topic. They will discuss the findings of each study, and also the strengths and constraints in each approach.

In paper [10] authors have used 34 features and collected URLs using scraping and other sources, used all possible algos. Their model detects the malicious urls from devices which can be handheld and only if those devices can support the Web browser. *Gap:* The work is done at only server side and not done at client side, also there is not no user experience yet. *Future work:* User experience as in the form of UI website or mobile app can be done in future.

[11] This paper is the further improvement of previously done work and demonstrates the previous approach proves ineffective when applied to the logs using actual proxy (a server between client and server that client wants to use) server due to the issue of imbalance. To address this, a new method is proposed, which follows a similar approach of extracting paragraphs from unknown proxy logs. In this method, the top n significant words are being extracted from the paragraphs obtained during the training process. These paragraphs are then converted into feature vectors using a trained Doc2vec model. *Gap:* It should be noted that the evaluation of this method has thus far been limited to data sets created by combining malign pcap files which store network packets with fine proxy logs. Its performance on different types of proxy logs remains to be assessed. *Future work:* This method can be used in the future for analyzing other proxy logs.

In [12] they have made a website using java, with fields of details about malicious URLs which will be entered by users. The ada boost model trained and will detect the URL as normal or malicious. *Gap:* no calculations have been added in this paper to prove what is done in the form of accuracy, precision metrics.

Also there is a list Ref. 2 of mentioned tools in few papers which might be helpful in making decision of usage of further findings by new or different enhanced libraries or tools. In table Ref. 4 there is a very comprehensive view of classifiers used. This can give a quick idea for which classifiers are common to use or may be which are left and can be included in upcoming work. And also that if the accuracy is leading in some results then which classifiers proved suitable for that type of dataset. Whereas Table Ref. 6 gives an idea of feature count that made much important role in preprocessing part.

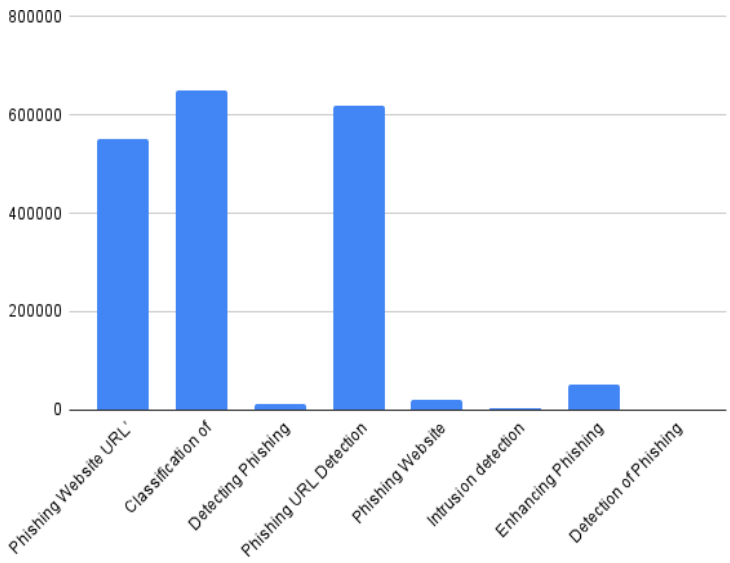

Fig. 3: Dataset Bar Chart

Fig. 4: Classifiers used

Title	Classifiers												
	naïve Bayes (NB)	decision tree (DT)	gradient boosted trees (GBT)	generalized linear model (GLM)	logistic regression (LR)	deep learning (DL)	random forest (RF)	Support Vector Machine (SVM)	AdaBoost algorithm	KNN	ANN	ExtraTree	XGBoost
Phishing Website URL's Detection Using NLP and Machine Learning Techniques	✓	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗
Classification of Malicious URLs Using Machine Learning	✗	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗	✗
Detecting Phishing Domains Using Machine Learning	✗	✓	✗	✗	✗	✗	✓	✓	✗	✗	✓	✗	✗
Phishing URL Detection Using Machine Learning	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✓	✗
Phishing Website Classification using Machine Learning with Different Datasets	✗	✓	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✗
Intrusion detection based on phishing detection with machine learning	✗	✓	✗	✗	✗	✗	✓	✗	✗	✓	✓	✗	✗
Enhancing Phishing URL Detection Accuracy in Software-Dened Networks (SDNs) through Feature Selection and Machine Learning Techniques	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Detection of Phishing Sites Using Machine Learning Techniques	✗	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✓

Fig. 5: Applied Feature Engineering

Title	Feature Count	Applied Feature Engineering	Feature Extraction				
			Domain based	Host based	Reputation based	Lexical	Content based
Phishing Website URL's Detection Using NLP and Machine Learning Techniques	-	✓	✗	✓	✓	✓	✓
Classification of Malicious URLs Using Machine Learning	16	✓	✗	✓	✗	✓	✗
Detecting Phishing Domains Using Machine Learning	30	✓	✗	✗	✗	✗	✗
Phishing URL Detection Using Machine Learning	7	✓	✓	✗	✓	✓	✗
Phishing Website Classification using Machine Learning with Different Datasets	-	✓	✗	✗	✗	✓	✗
Intrusion detection based on phishing detection with machine learning	5	✓	✗	✗	✗	✗	✗
Enhancing Phishing URL Detection Accuracy in Software-Dened Networks (SDNs) through Feature Selection and Machine Learning Techniques	30	✓	✓	✗	✓	✗	✓
Detection of Phishing Sites Using Machine Learning Techniques	9	✓	✓	✓	✓	✓	✓

The Tabular comparison is for papers in 2023 and 2024 which are [1], [2], [3], [4], [5], [6], [7], [8].

Fig. 6: Literature Review

Title	Year	Paper Type	Dataset Type							Dataset						Feature Count	Applied Feature Engineering	Feature Extraction						Tools/Libraries	Classifiers																Outcomes				Gap	Future Work
			Domain names	Proxy Server Logs	Pcap Files	Login URLs	Web App URLs	IP address	resource path	malicious		normal		total				Domain -based	Host-based	Reputation-based	Lexical	Content -based		naive Bayes (NB)	decision tree (DT)	gradient boosted trees (GBT)	generalized linear model (GLM)	logistic regression (LR)	deep learning (DL)	random forest (RF)	Support Vector Machine (SVM)	AdaBoost algorithm	KNN	ANN	ExtraTree	XGBoost	accuracy (%)	percision (%)	recall (%)	F measure					
										number	source	number	source	number	source																										timeline analysis	other dataset			
Leaving all proxy server logs to paragraph vector. J. Inf. Process	2018	journal of information processing	✗	✓	✓	✗	✗	✗	✗	7629	Exploit Kits (2014 to 2017). downloaded from MALWARE-TRAFFIC-ANALYSIS.NET	608796	Actual proxy logs (2016 to 2017) from a campus network	616425		10	✓	✗	✗	✗	✗	✗	✓	gensim-1.0.1, scikit-learn0.18.0 and chainer-1.23	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗			97	96	not yet been applied to actual proxy server logs	apply method to actual proxy server logs, adjust the size of a paragraph. use other NLP techniques to summarize a paragraph
Malicious Domain Names Detection Algorithm Based on Lexical Analysis and Feature	2019	Journal of electrical and electronics	✓	✗	✗	✗	✗	✗	✗	11000	Phishing Tank Newgoz Shiotob	13000	Alexa Anquan Organization	24000		.	✗	✓	✗	✗	✗	✗	Four metrics of LA-FQ (Latency-Aware Fair Queuing)	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	94.58					proposed method of using lexical features and characters distribution is not comprehensive and cannot detect all malicious domain names on the Internet, if the malicious domain	will be based on both further refinement of the methods, and a more sophisticated analysis using more substantial data sets
Malicious URL Detection based on Machine Learning	2020	International Journal of Advanced Computer Science and Applications	✗	✗	✗	✗	✓	✗	✗	70000	Phishtank URLhaus	400000	Alexa	470000		54	✓	✗	✓	✗	✓	✓		✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	91.11	91.43		89.72		Readers are encouraged to implement some other algorithms such as Naïve Bayes, Decision trees, k-nearest neighbors, neural networks, etc	The results of this research can be applied and implemented in information security technologies in information security systems
Phishing Website URL's Detection Using NLP and Machine Learning Techniques	2023	Journal on Artificial Intelligents	✗	✗	✗	✗	✓	✗	✗	150000		400000		550000		.	✓	✗	✓	✓	✓	✓	Python 3.8, scikit-learn 0.24, TensorFlow 2.4, NLTK 3.5 Development tool—DataIore JetBrains	✓	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	90 NB,LSVC,LR	94 RF		74 NB	lacks a clear strategy for adapting to dynamic phishing techniques.	focus on adaptive models and explore advanced techniques like deep learning for enhanced cybersecurity.
Classification of Malicious URLs Using Machine Learning	2023	Sensors - scientific journal - by MDPI	✗	✗	✗	✗	✓	✗	✗					650000	Kaggle Repo	16	✓	✗	✓	✗	✓	✗	Matlab 2022	✗	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗	✗	-	93.19 RF	91.19 RF / 92.18 RF		limiting insight into alternative approaches for superior malicious URL identification	Model Diversification: Explore advanced models (neural networks, NB, XGB, LGB) to identify superior approaches for malicious URL detection. Expanded URL Categories: Include additional categories (redirect URLs, scam URLs, clickbait URLs, drive-by downloads) for a more comprehensive understanding of cyber threats. Dataset Enrichment: Integrate larger and diverse datasets to strengthen model performance and improve practical applicability in malicious URL detection.
Detecting Phishing Domains Using Machine Learning	2023	Applied sciences - by MDPI	✗	✗	✗	✗	✓	✗	✗					11055	University of California, Irvine UCI Phishing Websites	30	✓	✗	✗	✗	✗	✗		✗	✓	✗	✗	✗	✗	✓	✓	✗	✗	✓	✗	✗	97.3 RF	96.9 RF		98.62 RF / 97.6 RF	After Use Normalization 1 % increase observed	examining more machine learning algorithm techniques for phishing domains
Phishing URL Detection Using Machine Learning	2024	IRE Journals	✓	✗	✗	✗	✓	✓	✗					620000	kaggle, phishtank, Aalto University URlSet	7	✓	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗	✗	✓	✗	✗	80.67 ExtraTree > RF					enhancing phishing detection by dissecting URL structures, analyzing user behavior, scrutinizing web/email content for anomalies, studying evasion tactics, implementing adaptive systems, and improving model transparency for user trust
Phishing Website Classification using Machine Learning with Different Datasets	2024	International Journal of Computing and Digital Systems	✓	✗	✗	✗	✓	✗	✗					20000	URL 2016 from Canadian Institute for Cybersecurity URL 2021 Mendeley Data		✓	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✗	✗	90.37 for 2021 99 for 2016					aims to advance phishing detection by developing a model that can navigate complex URLs and detect subtle intricacies. Integration of supplementary techniques like URL, HTML Encoding, WHOIS method, Tiny URL approach, and a sophisticated voting technique is expected to enhance detection accuracy and robustness.
Intrusion detection based on phishing detection with machine learning	2024	journal of measurement-sensors	✗	✗	✗	✗	✓	✗	✗					3000		5	✓	✗	✗	✗	✗	✗	Python Weka	✗	✓	✗	✗	✗	✗	✓	✗	✗	✓	✓	✗	✗	97.6 for proposed method					
Enhancing Phishing URL Detection Accuracy in Software-Defined Networks (SDNs) through Feature Selection and Machine Learning Techniques	2024	research square	✓	✗	✗	✗	✓	✗	✗	12000	phishtank	39000	5000best	51100		30	✓	✓	✗	✓	✗	✓	Open Network Operating System controller, Mininet tool	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	99.65	99.52		99.38/99.36		
Detection of Phishing Sites Using Machine Learning Techniques	2024	International Research Journal on Advanced Engineering Hub	✓	✗	✗	✗	✗	✓	✗							9	✓	✓	✓	✓	✓	✓		✗	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✓	87 XGBoost					will explore the integration of deep learning for enhanced feature extraction from webpage code and text. Additionally, we plan to create a browser plugin to seamlessly incorporate our approach into web browsers.

The detailed literature review in tabular form combined and additional information has also been added like if new thing introduced in any paper etc. The papers are in sequence from 1 to 11 as in reference.

III. METHODOLOGY

The machine learning approach for detecting malicious URLs typically involves two noticeable phases: the training phase and the detection phase. While in the training phase, a large data set of URLs is used to train a machine learning model to recognize patterns and features that distinguish between malicious and benign URLs. This involves two key steps: feature extraction and labeling. Feature extraction involves identifying specific characteristics, or features, of the URLs that could be useful in distinguishing between malicious and benign URLs. Examples of features include lexical features (e.g., identification of particular words or characters within the URL), content-based features (e.g., the presence of specific HTML tags), and domain-based features (e.g., the length or age of the domain). Once the features are extracted, each URL in the dataset is labeled as either malicious or benign based on whether it has been previously identified as such. The labeled dataset is then used to train a ML algorithm to check the patterns and also the features that distinguish between malign and benign URLs. Once the algorithm is trained, it moves on to the detection phase. In this phase, the algorithm takes in a new URL and extracts the same features that were used in the training phase. It then uses these features to classify the URL as either malign or benign. If the URL is classified as malicious, appropriate action can be taken, such as blocking access to the URL. On the other hand, if the URL is classified as benign, access can be allowed.

Overall, the feature extraction and classification process play an important role in the training phase and also in the detection phase of machine-learning based malign URL detection. By using a variety of features to accurately classify URLs, these methods provide an efficient and effective way to identify potentially harmful URLs and protect users from online threats.

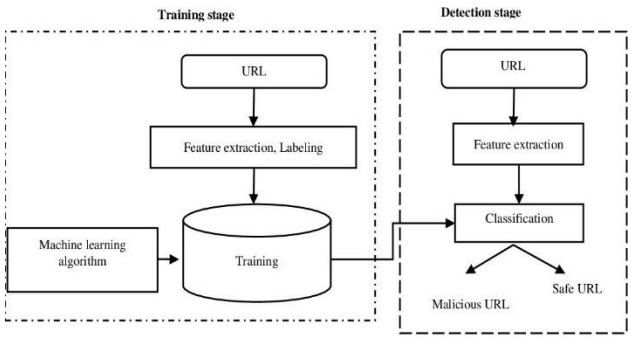

Fig. 7: Approach for Malign URL Detection based on Machine Learning

Title	Outcomes		
	accuracy (%)	percision (%)	recall/F measure (%)
Phishing Website URL's Detection Using NLP and Machine Learning Techniques	90 for NB,LSVC,LR	94 for RF	74 for NB
Classification of Malicious URLs Using Machine Learning	-	93.19 for RF	91.19 for RF/ 92.18 for RF
Detecting Phishing Domains Using Machine Learning	97.3 for RF	96.9 for RF	98.62 for RF/ 97.6 for RF
Phishing URL Detection Using Machine Learning	80.67 for ExtraTree	-	-
Phishing Website Classification using Machine Learning with Different Datasets	90.37 for 2021, 99 for 2016	-	-
Intrusion detection based on phishing detection with machine learning	97.6 for proposed method	-	-
Enhancing Phishing URL Detection Accuracy in Software-Dened Networks (SDNs) through Feature Selection and Machine Learning Techniques	99.65	99.52	99.38/99.36
Detection of Phishing Sites Using Machine Learning Techniques	87 for XGBoost	-	-

Fig. 8: Results

IV. CONCLUSION

This paper reviewed different research papers that focused on the topic of malign URL detection with the usage of Machine Learning approaches basically with selection of various features and dataset types. These papers shed light on various aspects of software development and provided valuable insights into the importance of measuring and analyzing different metrics. One of the papers, in particular, has been used to emphasize the relevance of metrics in software testing. It was shown that by measuring the attributes, software developers and researchers can assess the effectiveness of their testing strategies and identify areas where improvements are needed. Overall, this paper highlights the importance of using metrics to guide software development processes and the potential benefits that can be achieved by doing so.

From the pie chart Figure 9 for the feature count included in different papers discussed above it can be seen that the one with high percentage of count is having good results as compared to the others, see in results table Ref. 8. So, by the fact that with increasing the number of features, one can capture a broader range of information, allowing the model to consider more factors when making predictions. This can lead to a more comprehensive understanding of the underlying patterns in the data. However, it's essential to strike a balance and avoid over fitting by selecting relevant features and applying appropriate regularization techniques. There can be more improvements done by including other methods in feature extraction like character based, entropy based, path

based, redirect based or by combing all of them with diversity of dataset.

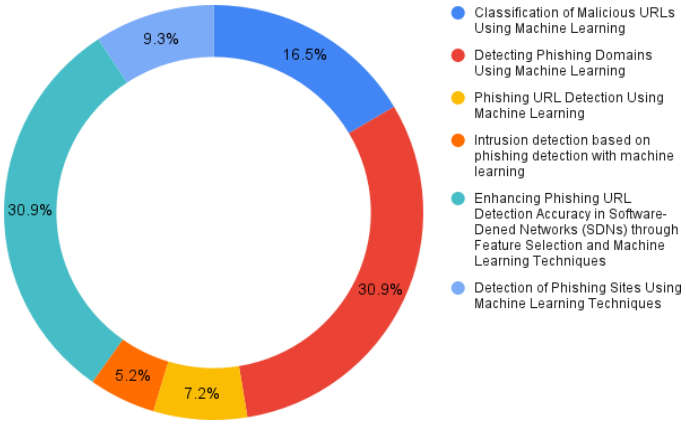

Fig. 9: Feature Count

REFERENCES

[1] C. D. Xuan, H. D. Nguyen, and T. V. Nikolaevich, "Malicious url detection based on machine learning," International Journal of Advanced Computer Science and Applications, vol. 11, no. 1, 2020. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2020.0110119

[2] T. Li, G. Kou, and Y. Peng, "Improving malicious urls detection via feature engineering: Linear and nonlinear space transformation methods," Information Systems, vol. 91, p. 101494, 07 2020.

[3] M. Sánchez-Paniagua, E. F. Fernández, E. Alegre, W. Al-Nabki, and González-Castro, "Phishing url detection: A real-case scenario through login urls," IEEE Access, vol. 10, pp. 42 949–42 960, 2022.

[4] S. H. Ahammad, S. D. Kale, G. D. Upadhye, S. D. Pande, E. V. Babu, A. V. Dhumane, and M. D. K. J. Bahadur, "Phishing url detection using machine learning methods," Advances in Engineering Software, vol. 173, p. 103288, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0965997822001892

[5] H. Bouijij and A. Berqia, "Phishing website classification using machine learning with different datasets," International Journal of Cyber Defense and Digital Forensics, vol. 9, no. 1, 2024. [Online]. Available: http://dx.doi.org/10.12785/ijcds/1501115

[6] R. Jayaraj, A. Pushpalatha, K. Sangeetha, T. Kamaleshwar, S. Udhaya Shree, and D. Damodaran, "Intrusion detection based on phishing detection with machine learning," Measurement: Sensors, vol. 31, p. 101003, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2665917423003392

[7] U. RUBY and G. C. C. J, "Enhancing phishing url detection accuracy in software-defined networks (sdns) through feature selection and machine learning techniques," 2024. [Online]. Available: https://doi.org/10.21203/rs.3.rs-3955168/v1

[8] A. Mahmood, V. Pandey, R. Raj, G. Shankar, and Mishra, "Detection of phishing sites using machine learning techniques," International Research Journal on Advanced Engineering Hub (IRJAEH), 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268493894

[9] M. Mimura, "Adjusting lexical features of actual proxy logs for intrusion detection," Journal of Information Security and Applications, vol. 50, no. C, pp. 128 990–128 999, 2020. [Online]. Available: https://doi.org/10.1016/j.jisa.2019.102408

[10] S. Anwar, F. Al-Obeidat, A. Tubaishat, S. Din, A. Ahmad, F. A. Khan, G. Jeon, and J. Loo, "Countering malicious urls in internet of things using a knowledge-based approach and a simulated expert," IEEE Internet of Things Journal, vol. 7, no. 5, pp. 4497–4504, 2020.

[11] M. Mimura, "Adjusting lexical features of actual proxy logs for intrusion detection," Journal of Information Security and Applications, vol. 50, p. 102408, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214212619305101

[12] F. Khan, J. Ahamed, S. Kadry, and L. K. Ramasamy, "Detecting malicious urls using binary classification through adaboost algorithm," International Journal of Electrical and Computer Engineering, vol. 10, no. 1, 2020. [Online]. Available: http://doi.org/10.11591/ijece.v10i1.pp997-1005.