e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

ENHANCING WEB SECURITY THROUGH AN OPTIMIZED LSD MODEL FOR ACCURATE AND EFFICIENT SUSPICIOUS WEB SITE DETECTION

D Lokesh[*1], Adeppa Tharun Kumar[*2], Abburi Karthik[*3], G Karthikeya Bhargava Sharm[*4], Dr. R. Karunia Krishnapriya[*5], Mr. N. Vijaya Kumar[*6], Mr. Pandreti Praveen[*7], Mr. V Shaik Mohammad Shahil[*8]

[*1,2,3,4]UG Scholar, Sreenivasa Institute Of Technology And Management Studies, Chittoor, India.

[*5]Associate Professor, Sreenivasa Institute Of Technology And Management Studies, Chittoor, India.

[*6,7,8]Assistant Professor, Sreenivasa Institute Of Technology And Management Studies, Chittoor, India.

DOI : https://www.doi.org/10.56726/IRJMETS71616

ABSTRACT

Phishing is a common cyberthreat that uses phony websites to trick people into disclosing private information. This study introduces a phishing detection system that analyses URLs and categorizes them as malicious or authentic using a hybrid machine learning technique. In addition to content-based and behavioural indications, the system extracts important URL-based parameters like length, domain age, presence of special characters, and entropy. To improve detection accuracy, a variety of supervised machine learning methods are used, including Random Forest, SVM, and neural networks. Prior to engaging with them, the suggested high precision and recall. When compared to conventional blacklist-based techniques the approach greatly improves phishing detection and raises awareness of cybersecurity

Keywords: Phishing Detection, Cybersecurity, Machine Learning, Hybrid Model, Url Analysis, Feature Extraction, Supervised Learning, Web Security, Random Forest, Neural Networks, Cyber Threats.

I. INTRODUCTION

Phishing attacks have grown to be a serious cybersecurity risk since they deceive users into divulging private information including login passwords, bank account information, and personal information. Because fraudulent websites are often quite similar to authentic ones, it is hard for consumers to tell the difference between the two. Because they rely on pre-existing threat databases, traditional phishing detection techniques like rule-based and blacklist-based approaches lack the ability to identify recently launched phishing sites. This work proposes a phishing websites Detection System based on the LSD hybrid Model Which stands for Decision tree, Support Vector Machine and logistic Regression. The model combines several machine learning techniques to improve phishing detections precision and effectiveness. The suggested algorithm can identify patterns liked to harmful websites because it is trained on a dataset that includes both genuine and phishing URLs. The LSD hybrid Model enhances detection accuracy and lowers false positives by integrating SVM for boundary classification, Detection, giving visitors a simple way to confirm the legitimacy of websites before visiting them. By presenting a reliable and expandable phishing detection method, this study advances the subject of cybersecurity. People business, and security experts can all benefit from the study's conclusions to improve online safety and successfully counteract phishing attacks.

II. LITERATURE SURVEY ON WORKLOAD

PREDICTION

In the world of cybersecurity, Phishing attacks have been thoroughly examined, and a number of detecting techniques, which depend on heuristic and blacklist -based techniques, have trouble in identifying new phishing websites. The necessity for a hybrid machine learning-based method such as the LSD hybrid model is highlighted in this section, which summarizes previous research in phishing detection.

1.Detection of phishing with Blacklists: blacklists that contain known phishing website URLs are the foundation of conventional phishing detection known phishing website URLs are the foundation of conventional phishing detection methods. Two popular blacklist-based programs that stop users from visiting harmful websites are google safe browsing and Microsoft smart screen. However these approaches have drawbacks such delayed updates, Expensive maintenance costs, and an inability to identify recently developing phishing sites. According to research real-time detection is difficulty because over 50% of phishing websites go unnoticed in blacklists for the first few hours of their existence.

2.Heuristic-Based Methodologies: To find questionable websites, heuristic-based phishing detection systems examine domain information, URL traits and websites structures. Classification is done using attributes such URL length number of subdomains, presence of special characters and SSL certificate status. Because legitimate websites are dynamic, heuristic approaches frequently result in false positives even if they are better at detecting phishing than blacklists.

3.Machine Learning-Oriented Strategies: Because ML approaches can identify trends in URLs and website content, they have become popular for detection phishing attempts. To categorize phishing websites, neural networks, decision trees, random forests and support vector machines have been used These models information. The computational complexity, high false positive rates, and poor generalization capacity are some of the drawback of single-machine learning models.

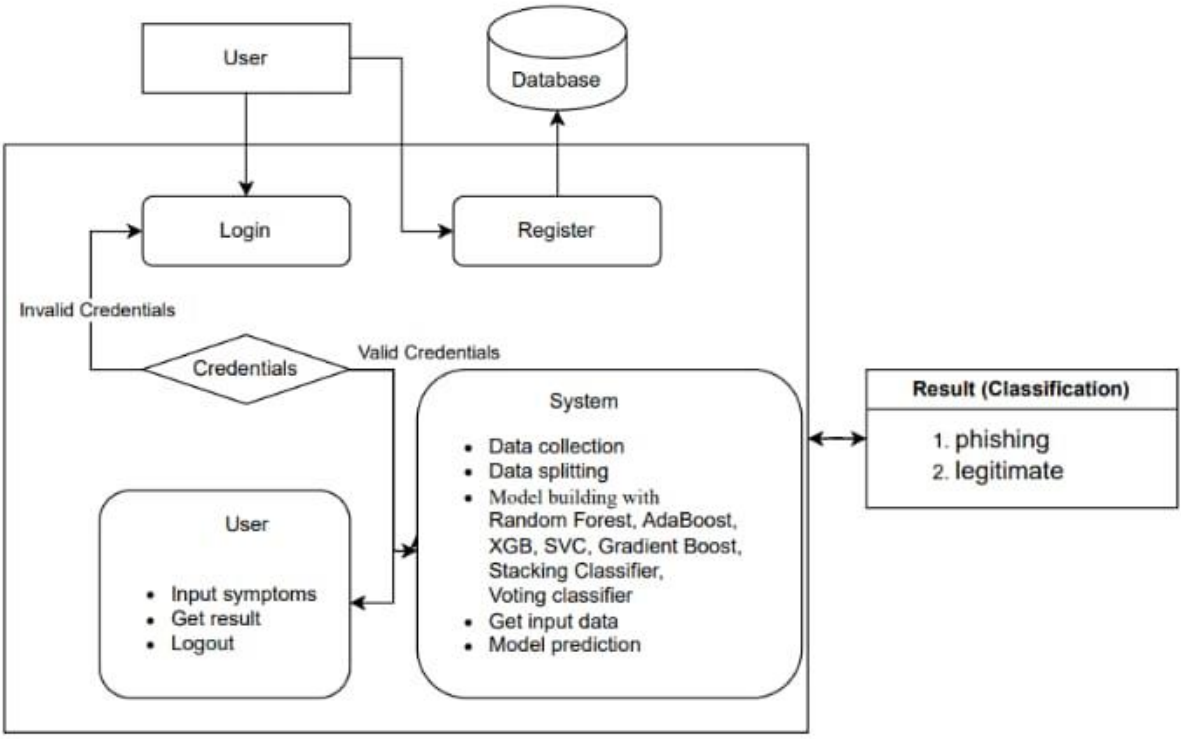
4.Phishing detection using hybrid machine learning models; The usefulness of hybrid machine learning models in phishing detection has been demonstrated by recent studies. A hybrid strategy that combined Random Forest and SVM was presented by Gupta et al. (2023) and produced a detection accuracy of more than 95%. Likewise, Kumar et al. (2024) created a hybrid model that combined decision trees with neural networks, which resulted in a20% decrease in false positives. These experiments show that phishing detection performances is improved by integrating many machines learning techniques.

5.LSD Hybrid Model for identifying Phishing websites: This study suggests a Phishing website detection system utilizing the LSD Hybrid Model in light of the shortcomings noted in earlier research. To increase accuracy and reduce false positives

III. METHODOLOGY

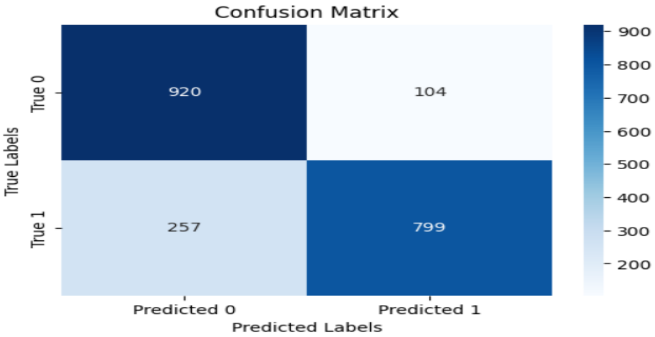
This Phishing websites detection using machine learning project's methodology is broken down into a number of crucial components, all of which are aimed at creating a reliable system for detecting websites using different machine learning algorithms. An outline of the process is provided below

ARCHITECTURE DIAGRAM



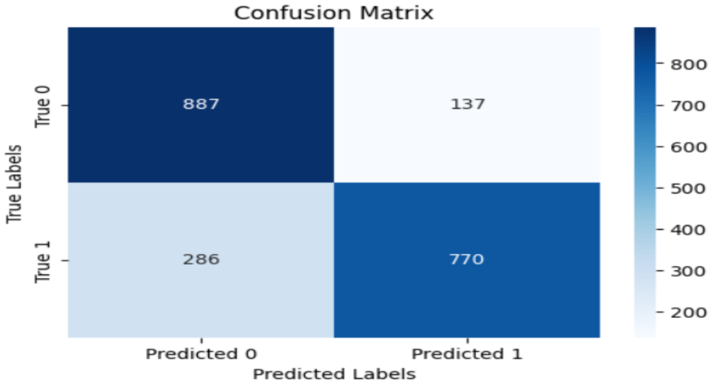
1.Random Forest classifier: Gather the dataset of trustworthy and phishing websites, then preprocess it by handling missing values and eliminating superfluous columns. Extract characteristic including URL structure, domain information, and website content that are important for spotting phishing website. Utilize an ensemble of decision trees produced by the Random Forest methods. A random selection of data features is used to train each tree, and the majority vote from all trees is used to determine the final prediction.

- Results:
- Accuracy:83%
- Precision:0.78(for class 0),0.88(for class 1)
- Recall: 0.90(for class 0),0.76(for class 1)
- F1-Score:0.84(for class 0),0.82(for class 1)



2.Adaboost classifier: An ensemble learning approach called Adaptive Boosting improves the performance of weak classifiers by training several models in succession and modifying their weights in responses to cases that are incorrectly identified. By concentrating more on challenging-to-classify phishing websites, AdaBoost enhances classification accuracy in phishing website detection.

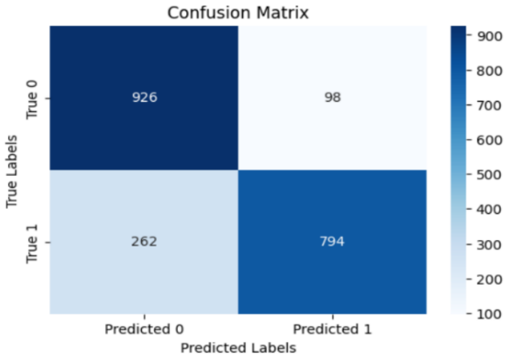
- Results:
- Accuracy: 80%
- Precision:0.76(for class 0),0.85(for class 1)
- Recall:0.87(for class 0),0.73(for class 1)
- F1-Score:0.81 (for class 0),0.78(For class 1)



3.XGBoost Classifier: A sophisticated ensemble machine learning technique based on gradient boosting, Extreme Gradient Boosting (XG Boost) is tuned for accuracy and speed. Its capacity to manage huge datasets, lessen overfitting, and enhance classification performance makes it a popular choice for phishing website identification. The way XG Boost operates is by successively building several decision trees, each of which fixes the mistakes of the mistakes of the one before it. It improves prediction accuracy and avoids overfitting by using regularization approaches (L1 and L2).

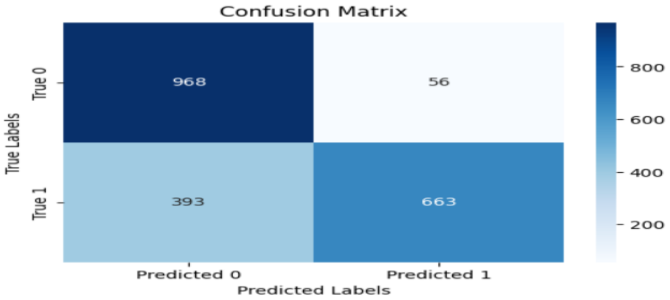
- Results:
- Accuracy: 83%
- Precision: 0.78(for class 0),0.89 (for class 1)
- Recall:0.90(for class 0),0.89(for class 1)

- F1-Score: 0.84(for class 0),0.82 (for class 1)



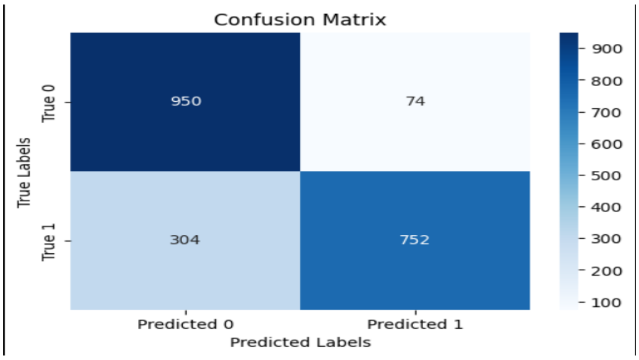
4.Support Vector Machine: The robust supervised learning method support Vector machine is used for classification tasks, such as identifying phishing websites. In order to improve generality, it finds the best hyperplane to divide phishing and trustworthy websites with the greatest margin. Even with tiny datasets, SVM can function effectively and is especially good at handling high-dimensional data. Through the analysis of numerous URL-based, domain-based, and content-based characteristics, SVM determine whether a given URL is Phishing or authentic.

- Results:
- Accuracy:78%
- Precision: 0.71(for class 0),0.92 (for class 1)
- Recall:0.95(for class 0), 0.63(for class 1)
- F1-Score:0.81(for class 0),0.75(for class 1)



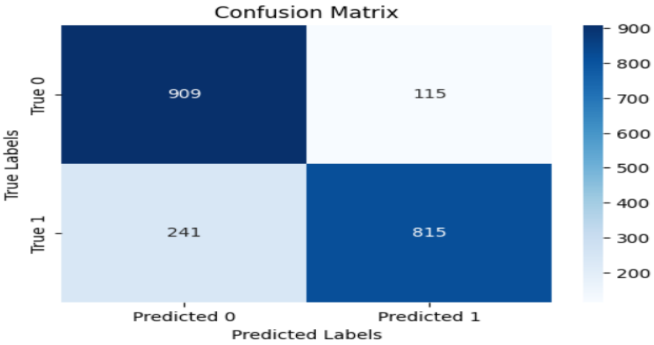
5.Gradient Boosting Classifier: A potent ensemble learning method for identifying fraudulent websites is gradient boosting. It Construct several weak learners in a sequential fashion, with each new tree concentrating on fixing the errors of the ones that came before it. Gradient Boosting build a robust classifier that can accurately differentiate between phishing and trustworthy websites by iteratively reducing errors.

- Results:
- Accuracy: 82%
- Precision:0.76 (for class 0),0.91 (for class 1)
- Recall:0.93(for class 0),0.71 (for class 1)
- F1-Score:0.83 (for class 0), 0.80(for class 1)



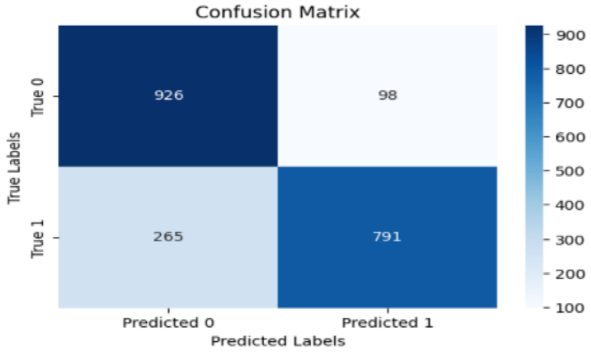
6. Stacking Classifier: To increase classification accuracy, the Stacking Classifier, a sophisticated ensemble learning method, mixes several base models. Stacking learns from the prediction of several base classifiers using a meta-model, in contrast to bagging and boosting

- Results:
- Accuracy: 83%
- Precision: 0.79(for class 0),0.88(for class 1)
- Recall:0.89(for class 0), 0.77(for class 1)
- F1-Score: 0.84(for class 0), 0.82(for class 1)



7.Voting Classifier: To increase classification accuracy, the Voting Classifier is an ensemble learning method that blends several machine learning models. To get a final choice, it combines the prediction of multiple classifiers and chooses either the average projected probability or the majority vote.

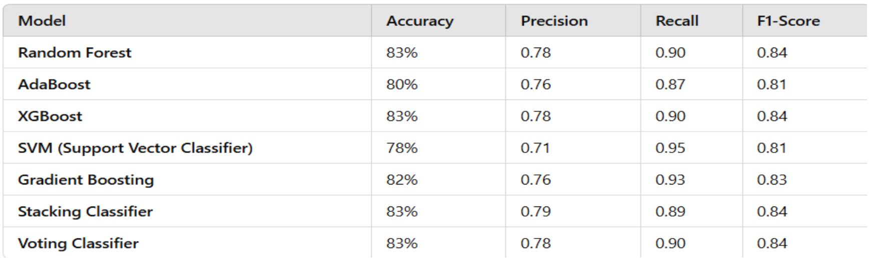
- Results:
- Accuracy: 83%
- Precision: 0.79(for class 0),0.88(for class 1)
- Recall: 0.89(for class 0), 0.82 (for class 1)
- F1-Score: 0.84(for class 0),0.82 (for class 1)



IV. IMPLEMENTATION AND RESULTS

Several machine learning models were trained and tested on the phishing detection dataset, and their performance was accessed using important metrics like F1-score , accuracy ,precision and recall .The effectiveness of each model in differentiating between phishing and authentic websites is evaluated with the use of these criteria

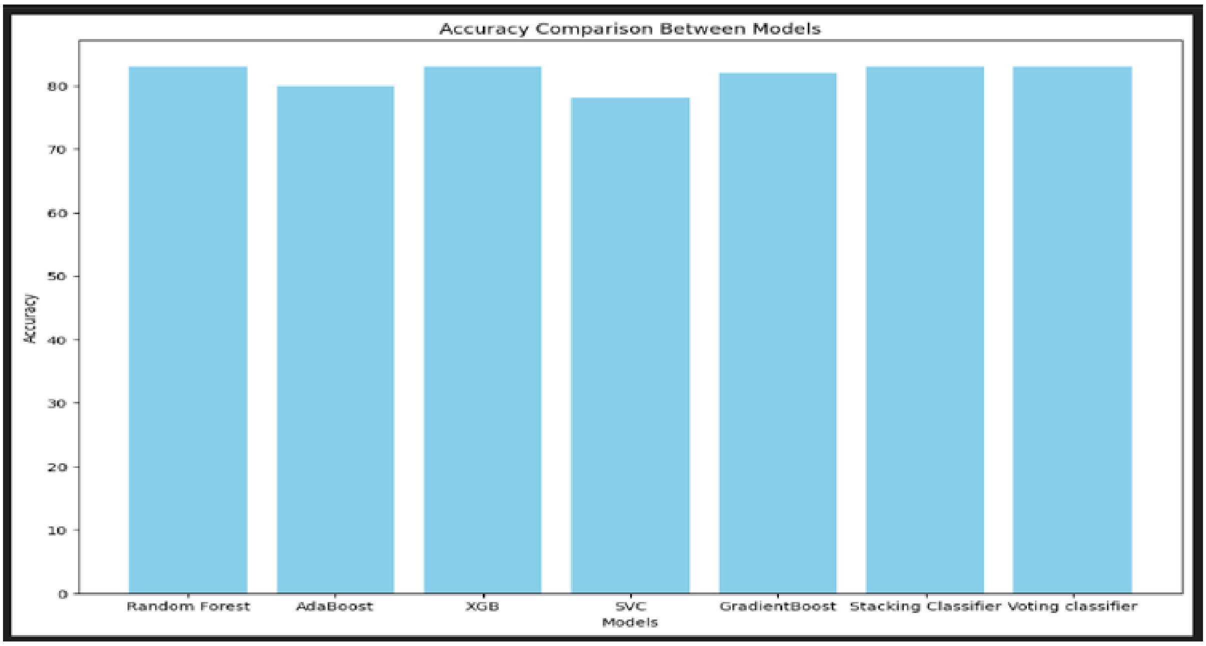
Accuracy Comparison of Different Models



Model	Accuracy	Precision	Recall	F1-Score
Random Forest	83%	0.78	0.90	0.84
AdaBoost	80%	0.76	0.87	0.81
XGBoost	83%	0.78	0.90	0.84
SVM (Support Vector Classifier)	78%	0.71	0.95	0.81
Gradient Boosting	82%	0.76	0.93	0.83
Stacking Classifier	83%	0.79	0.89	0.84
Voting Classifier	83%	0.78	0.90	0.84

The accuracy scores of each machine learning model were compared in order to access their efficiency. Since accuracy shows how well a model can distinguish between phishing and legal websites, it is an essential metric in phishing detection

Accuracy Comparison Graph



V. CONCLUSION

Machine learning has been shown to be a successful strategy in the fight against cyberthreats for detecting phishing websites. Multiple machine learning methods, such as Random Forest, XG Boost, AdaBoost, Gradient Boosting, SVM, Stacking and Voting Classifiers, were used in this work to categorize websites as either real or phishing. According to the results ensemble models that outperformed conventional classification techniques, including Random Forest, XG Boost, Stacking and Voting Classifier, attained the greatest accuracy of 83% By examining attributes based on URLs, domains, and content, the LSD Hybrid Model (Logistic Regression, SVM, and Decision Tree) offered a productive method of identifying phishing websites. An online platform added a even more functionality to the system by enabling users to detect phishing attempts in real time.

ACKNOWLEDGMENT

we want to sincerely thank our university, instructors and mentors for their unwavering support and direction our study their encouraging words and insightful advice were really helpful in constructing this research. We also give credit to academic journals and the opensource community for providing the datasets and expertise that were crucial to our end. We are very grateful to Kaggle for sharing datasets of authentic and phishing websites, which made it easier to train the machine learning model. Finally, we would like to thank our teams for their commitment, cooperation, and diligence in putting this phishing detection system into placed and accessing its effectiveness. Without their combined efforts, this research would not have been feasible

VI. REFERENCES

[1] L. Jovanovic et al., "Improving Phishing Website Detection using a Hybrid Two-level Framework for Feature Selection and XG Boost Tuning," in Journal of Web Engineering, vol. 22, no. 3, pp. 543-574, May 2023.

[2] S. Ahmad et al., "Across the Spectrum In-Depth Review AI-Based Models for Phishing Detection," in IEEE Open Journal of the Communications Society, 2022.

[3] L. R. Kalabarige, R. S. Rao, A. R. Pais and L. A. Gabralla, "A Boosting-Based Hybrid Feature Selection and Multi-Layer Stacked Ensemble Learning Model to Detect Phishing Websites," in IEEE Access, vol. 11, pp. 71180-71193, 2023.

[4] Rishikesh Mahajan, Irfan Siddavatam, "Phishing Website Detection using Machine Learning Algorithms", researchgates, 2018.

[5] Y. A. Alsariera, V. E. Adeyemo, A. O. Balogun and A. K. Alazzawi, "AI Meta-Learners and Extra-Trees Algorithm for the Detection of Phishing Websites," in IEEE Access, vol. 8, pp. 142532-142542, 2020.

[6] J. Shad and S. Sharma, "A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology," pp. 425–430, 2018

[7] Y. Sönmez, T. Tuncer, H. Gökal, and E. Avci, "Phishing web sites features classification based on extreme learning machine," 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.

[8] T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018–Janua, pp. 300–301, 2018.

[9] M. Karabatak and T. Mustafa, "Performance comparison of classifiers on reduced phishing website dataset," 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.

[10] S. Parekh, D. Parikh, S. Kotak, and P. S. Sankhe, "A New Method for Detection of Phishing Websites: URL Detection," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, vol. 0, no. Icicct, pp. 949–952.

[11] K. Shima et al., "Classification of URL bitstreams using bag of bytes," in 2018 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), 2018, vol. 91, pp. 1–5.

[12] A. Vazhayil, R. Vinayakumar, and K. Soman, "Comparative Study of the Detection of Malicious URLs Using Shallow and Deep Networks," in 2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018, 2018, pp. 1– 6.

[13] W. Fadheel, M. Abusharkh, and I. Abdel-Qader, "On Feature Selection for the Prediction of Phishing Websites," 2017 IEEE 15th Intl Conf Dependable, Auton. Secur. Comput. 15th Intl Conf Pervasive Intell. Comput. 3rd Intl Conf Big Data Intell. Comput. Cyber Sci. Technol. Congr., pp. 871–876, 2017.

[14] X. Zhang, Y. Zeng, X. Jin, Z. Yan, and G. Geng, "Boosting the Phishing Detection Performance by Semantic Analysis," 2017.

[15] L. MacHado and J. Gadge, "Phishing Sites Detection Based on C4.5 Decision Tree Algorithm," in 2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA 2017, 2018, pp. 1–5