International Journal of Engineering and Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 15 Issue 4 April 2026 || PP 71-77

Phishing Detection via Machine Learning-Based URL Analysis: A Comprehensive Survey

Shivam Singh[1], Dilip Kumar[2]
[1]M.Tech Scholar [C.S.E] , HOD Dept. of C.S.E.
Vishveshwarya Group of Institutions, Gautam Buddh Nagar

Abstract
The proliferation of internet-based transactions across financial, professional, and personal domains has correspondingly amplified the scale and sophistication of cyberthreats. Among these, phishing attacks—which leverage deceptive URLs to redirect unsuspecting users to fraudulent websites—constitute one of the most pervasive and damaging forms of cybercrime. Unlike technically oriented exploits that target system vulnerabilities, phishing exploits human cognitive weaknesses, making purely technical defences insufficient. This survey provides a systematic review of machine learning (ML)-based approaches for phishing URL detection, examining benchmark datasets, feature extraction methodologies, algorithmic classifiers, and performance evaluation metrics. The findings reveal that ensemble methods, particularly Random Forest, consistently demonstrate superior classification accuracy across diverse experimental configurations. This work aims to serve as a consolidated reference for researchers and practitioners engaged in developing robust phishing detection frameworks.

Keywords—*phishing detection; URL feature extraction; machine learning classifiers; cybersecurity; random forest; ensemble methods.*

I. INTRODUCTION

The widespread digitization of everyday activities—accelerated dramatically by the COVID-19 pandemic in 2020—created fertile conditions for cybercriminals to exploit the expanded online attack surface. As individuals and organizations shifted routine transactions to digital platforms, malicious actors intensified phishing campaigns designed to harvest sensitive credentials, financial data, and personally identifiable information (PII).

Phishing is a form of social engineering attack in which adversaries craft deceptive web pages and distribute malicious URLs through email, messaging platforms, or compromised websites. Victims are lured into believing these pages are legitimate, ultimately surrendering confidential information. The hybrid nature of phishing—combining technical deception with psychological manipulation—distinguishes it from conventional cyberattacks and demands equally multifaceted countermeasures.

The scale of phishing activity is well-documented and alarming. According to the FBI Internet Crime Report 2020 [2], phishing incidents nearly doubled from 114,702 reported cases in 2019 to 241,342 in 2020. Concurrently, the Verizon 2020 Data Breach Investigations Report [3] identified phishing as a contributing factor in approximately 22% of all data breaches recorded that year. The Anti-Phishing Working Group (APWG) [1] further confirmed that attack volumes against financial institutions surged in the fourth quarter of 2020, while the healthcare and retail sectors experienced significant targeting in the context of pandemic-related public concern.

Global health crises such as COVID-19 have provided additional pretexts for phishing campaigns. The World Health Organization [4] documented numerous instances where attackers impersonated health authorities, distributed fraudulent vaccine registration links, and deployed spear-phishing emails that capitalized on public anxiety. These campaigns illustrate how phishing strategies adapt rapidly to exploit societal events.

Traditional countermeasures—such as blacklists, manual URL inspection, and rule-based systems—have proven inadequate against evolving, polymorphic phishing tactics. Machine learning offers a promising alternative by enabling automated, scalable, and adaptive classification of URLs based on statistically learned patterns. This survey reviews and synthesizes the current body of research on ML-based phishing URL detection, with particular emphasis on feature engineering strategies, algorithmic performance, and open research challenges.

The remainder of this paper is organized as follows: Section II provides background on phishing detection approaches. Section III presents a structured literature review. Section IV describes the datasets used in prior studies. Section V outlines feature extraction methodologies. Section VI details performance evaluation metrics. Section VII presents consolidated observations, and Section VIII concludes with directions for future research.

II. BACKGROUND

A. Phishing Detection: Problem Formulation

At its core, phishing detection is a binary classification problem: given an input URL, the goal is to determine whether it belongs to a legitimate domain or constitutes a phishing attempt. The challenge lies in the dynamic and often subtle nature of phishing URLs, which are engineered to mimic legitimate counterparts while harboring malicious intent. Effective detection requires analysis of multiple URL attributes—structural, lexical, and behavioral—to reliably distinguish between the two classes.

B. Existing Detection Paradigms

Several detection paradigms have been explored in the literature, each with distinct strengths and limitations:

List-Based Methods: Blacklist and whitelist approaches represent the simplest form of phishing defense. A URL is flagged if it appears on a known malicious URL registry or denied access if absent from a whitelist [5], [6]. While computationally efficient, these methods are fundamentally reactive and fail to detect newly registered phishing domains that have not yet been catalogued.

Histogram-Based Methods: This approach models the statistical distribution of character patterns in known phishing URLs and classifies incoming URLs based on their similarity to these learned distributions [7]. Performance degrades for phishing URLs that deliberately avoid common lexical patterns.

Visual Similarity-Based Methods: These techniques compare the rendered visual layout of a web page against legitimate reference pages using image processing algorithms [8]. While effective for detecting high-fidelity impersonations, the approach is computationally intensive and impractical for real-time deployment at scale.

Content-Based Methods: By analyzing text and HTML content on web pages, these methods extract semantic features—such as brand name occurrences and keyword weightings—to infer legitimacy [9], [10]. However, reliance on third-party services (e.g., search engines, DNS records) introduces latency and privacy concerns.

Fuzzy Rule-Based Methods: Expert-designed fuzzy logic systems assign soft membership to feature categories, enabling nuanced handling of ambiguous or borderline URLs [11]. Performance is sensitive to feature selection quality; extraneous features degrade classifier accuracy.

Machine Learning-Based Methods: Supervised learning approaches train statistical models on labeled datasets of benign and phishing URLs, enabling probabilistic classification of unseen samples. This paradigm has emerged as the dominant research focus due to its adaptability, scalability, and consistently high accuracy, forming the primary subject of this survey.

C. Common Machine Learning Algorithms

Algorithms commonly applied to phishing URL detection include Naïve Bayes (NB), Decision Trees (DT), Support Vector Machines (SVM), Logistic Regression (LR), K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN), Random Forest (RF), Rotation Forest (RoF), and gradient boosting variants such as XGBoost. Each algorithm presents a different trade-off between computational complexity, interpretability, and classification performance.

III. LITERATURE REVIEW

This section reviews key studies that have applied ML-based approaches to phishing URL detection, focusing on experimental design, algorithmic comparisons, and notable findings.

Singh and Siddavatam [12] evaluated three classifiers—Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM)—on a dataset comprising 17,058 legitimate URLs sourced from the Alexa directory and 19,653 phishing URLs from PhishTank, encompassing 16 discriminative features. The dataset was partitioned using three train-test split ratios (50:50, 70:30, and 90:10) to assess model sensitivity to training data volume. The RF classifier achieved the highest accuracy of 97.14% at the 90:10 split, with a notably low false-negative rate. The study concluded that accuracy improves monotonically with increased training data, supporting the value of large-scale dataset collection.

Kumar et al. [13] constructed a balanced URL dataset with equal proportions of phishing and legitimate instances, from which lexical and structural features were extracted. Five classifiers—Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbor (KNN)—were trained using a 70:30 split. The Gaussian Naïve Bayes classifier achieved the highest area under the ROC curve (AUC = 0.991), while RF attained the highest accuracy at 98.03%. The study highlights that data balancing is a critical preprocessing step that substantially improves classifier reliability.

Korkmaz et al. [14] conducted a multi-algorithm, multi-dataset evaluation using eight classifiers—XGBoost, RF, DT, Naïve Bayes (NB), SVM, KNN, Logistic Regression (LR), and ANN—across three independent datasets, extracting 48 features from a candidate pool of 58. RF consistently achieved the highest

accuracy, peaking at 94.59% on Dataset 1. ANN emerged as a strong secondary performer. LR, SVM, and NB demonstrated comparatively limited performance. The authors noted that both RF and ANN offer favorable trade-offs between training time and classification accuracy, recommending them for practical deployment.

Alam et al. [15] proposed a feature selection–augmented approach employing dimensionality reduction techniques—including Recursive Feature Elimination (RFE), Relief-F, Information Gain (IG), Gain Ratio (GR), and Principal Component Analysis (PCA)—prior to training RF and DT classifiers on the 32-feature Kaggle phishing dataset. The RF model achieved 97% accuracy while demonstrating improved generalizability through reduced overfitting and lower variance compared to DT. The study underscores the importance of feature selection in improving both accuracy and model efficiency.

Subasi et al. [16] evaluated six classifiers—ANN, KNN, SVM, C4.5 Decision Tree, Random Forest, and Rotation Forest (RoF)—on the UCI Machine Learning Repository phishing dataset containing 11,055 records and 31 features. Performance was assessed across accuracy, F-measure, and AUC metrics. RF outperformed all competing classifiers on all three dimensions, demonstrating superior robustness and speed. The study affirms the consistency of RF as a high-performing classifier for phishing detection tasks.

Table I. Summary of Literature Review on ML-Based Phishing Detection

Ref.	Algorithms Used	Dataset & Features	Key Findings	Best Accuracy
[14]	XGBoost, RF, DT, NB, SVM, KNN, LR, ANN	3 datasets; 48 of 58 features (URL structural, domain, path)	RF ranks highest on all three datasets; ANN is also reliable. LR, SVM, NB underperform. RF and ANN balance speed with accuracy.	94.59% (DS1) 90.50% (DS2) 91.26% (DS3)
[12]	DT, RF, SVM	36,711 URLs (Alexa + PhishTank); 16 features; 50:50, 70:30, 90:10 splits	RF yields best accuracy and lowest false-negative rate. Accuracy rises with larger training portions.	96.72% (50:50) 96.84% (70:30) 97.14% (90:10)
[13]	LR, NB, RF, DT, KNN	Balanced URL dataset; lexical & structural features; 70:30 split	RF achieves highest accuracy. Gaussian NB attains the highest AUC (0.991). Data balancing is critical.	RF: 98.03% Gaussian NB: 97.18% (AUC = 0.991)
[16]	ANN, KNN, SVM, C4.5 DT, RF, RoF	UCI dataset; 11,055 records; 31 features; evaluated on Accuracy, F-measure, AUC	RF surpasses all classifiers on accuracy, F-measure, and AUC. It is faster and more robust.	RF: 97.36%
[15]	RF, DT (with PCA, RFE, IG, GR, Relief-F)	Kaggle phishing dataset; 32 features; feature selection applied before classification	Feature selection with RF reduces overfitting and variance. PCA improves efficiency without sacrificing accuracy.	RF: 97.00%

IV. DATASETS

The quality and composition of training datasets are foundational to the performance of any ML-based phishing detection system. The phishing URL datasets used in the surveyed literature are primarily derived from two public repositories:

PhishTank (phishtank.org): A community-driven, crowdsourced database of verified phishing URLs, accessible via API. PhishTank is widely used in academic research due to its frequent updates and verifiability. Notably, it stores only URL strings without associated page content, making it suitable for URL-centric analyses [14].

OpenPhish: A commercial phishing intelligence feed that automatically identifies phishing URLs using proprietary detection technologies. It is adopted by major cybersecurity organizations including APWG, Mozilla, McAfee, and Symantec.

Legitimate (benign) URLs are typically sampled from the Alexa Top Sites directory or the Common Crawl web corpus, providing a representative distribution of genuine web domains.

Two structured datasets are particularly prominent in the phishing detection literature:

UCI Machine Learning Repository Dataset [16]: Contains 11,055 URL records characterized by 31 binary or ternary feature attributes. Feature values encode the presence or severity of each attribute, with ternary features distinguishing between low, moderate, and high risk levels. The class label assigns +1 to legitimate URLs and -1 to phishing URLs.

Kaggle Phishing Dataset [15]: A publicly available dataset comprising 32 discriminative features, widely used for benchmarking ML classifiers on phishing detection tasks.

Dataset imbalance—where phishing samples are underrepresented relative to legitimate URLs—is a recognized challenge that can skew classifier performance. Several studies [13] explicitly address this through data balancing techniques such as oversampling and stratified sampling.

V. FEATURE EXTRACTION

Feature engineering is central to the effectiveness of URL-based phishing detection. A standard URL is composed of several syntactic components—protocol scheme, subdomain, domain name, top-level domain (TLD), path, query parameters, and fragment identifier. Each component carries potential discriminative information that can be extracted and encoded as a numerical feature vector.

Following the taxonomy proposed by Mohammad et al. [18], URL features are broadly categorized into four groups:

Table II. Feature Categories and Representative Attributes (UCI Dataset, Set I)

Feature Category	ID	Feature Name
Address Bar Based	1	Having IP Address
	2	URL Length
	3	URL Shortening Service
	4	Having '@' Symbol
	5	Double Slash Redirect
	6	Prefix/Suffix in Domain
	7	Subdomain Count
	8	SSL Certificate State
	9	Domain Registration Length
	10	Favicon Source
	11	Non-Standard Port
	12	HTTPS Token in Domain
Abnormal Based	13	Request URL
	14	URL of Anchor Tag
	15	Links in Meta / Script / Link Tags
	16	Server Form Handler (SFH)
	17	Submitting Form to Email
	18	Abnormal URL Pattern
HTML / JavaScript Based	19	Redirect Count
	20	OnMouseOver Behavior
	21	Right-Click Disabled
	22	Popup Window Usage
	23	Iframe Embedding
Domain Based	24	Domain Age
	25	DNS Record Existence
	26	Web Traffic Ranking
	27	PageRank Score
	28	Google Indexing Status
	29	Inbound Link Count
	30	Statistical Report Listing
	31	Result (Class Label)

Table III. Extended URL Feature Set (48 Features) — Korkmaz et al. [14]

#	Feature Name	#	Feature Name	#	Feature Name
1	Words in URL	**2**	URL Path Segments	**3**	Digit Count
4	Ampersand Count	**5**	Sensitive Keywords	**6**	'?' Presence
7	Special Characters	**8**	Punctuation Count	**9**	Dots in Subdomain
10	TLD in Path	**11**	Subdomain Presence	**12**	Digits in Hostname

#	Feature Name	#	Feature Name	#	Feature Name
13	Dot Count	**14**	Words in Hostname	**15**	Hyphen in Path
16	'=' Presence	**17**	Underscore Count	**18**	Dots in Host
19	Dots in Path	**20**	Hyphen in Host	**21**	URL Without 'www'
22	Query Parameters	**23**	Character Repetition	**24**	HTTPS Protocol
25	Digits in Domain	**26**	IP Address in URL	**27**	Subdomain Count
28	'www'/'com' Presence	**29**	'@' Symbol	**30**	Hyphen in URL
31	URL Suffix	**32**	Redirect Flag	**33**	Path Length
34	Subdomain Length	**35**	Full URL Length	**36**	Domain Name Length
37	Longest Word	**38**	Parameter Count	**39**	Average Word Length
40	Shortest Word	**41**	Longest Word in Host	**42**	Hostname Length
43	URL/Path Ratio	**44**	Vowel/Consonant Ratio	**45**	Digit/Letter Ratio
46	Longest/Shortest Word Ratio	**47**	Std. Dev. of Word Lengths	**48**	Port Number

Address Bar-Based Features (Features 1–12): These attributes are derived directly from the URL string and require no external lookups. They include indicators such as the presence of an IP address in lieu of a domain name, anomalous URL length, use of URL shortening services, inclusion of special characters (e.g., '@', '--'), and the use of non-standard ports or HTTPS tokens within the domain name. These features are computationally inexpensive and highly discriminative.

Abnormal-Based Features (Features 13–18): These characteristics capture deviations from normal web behavior, including mismatches between the URL and linked resources (Request URL, anchor tags, tags), the server form handler (SFH) destination, and URL patterns that match known suspicious templates.

HTML and JavaScript-Based Features (Features 19–23): Extracted from parsed page content, these features capture client-side scripting behaviors commonly associated with phishing pages, such as excessive redirects, disabled right-click functionality, invisible iframes, and deceptive onMouseOver events.

Domain-Based Features (Features 24–30): Derived from DNS records and domain registration databases, these features capture the historical and reputational attributes of a domain, including its age, PageRank, indexing status in major search engines, and appearance in blacklist reports. While highly informative, these features introduce latency due to third-party lookups.

Korkmaz et al. [14] extended this taxonomy to 48 features by decomposing URL components into granular numerical attributes (Table III), enabling richer statistical modeling of URL characteristics without reliance on external services.

VI. PERFORMANCE EVALUATION METRICS

The performance of phishing detection classifiers is assessed through a standardized set of evaluation metrics derived from the confusion matrix, which tabulates classifier predictions against ground truth labels across four fundamental outcomes:

True Positive (TP): A phishing URL correctly classified as phishing. True Negative (TN): A legitimate URL correctly classified as legitimate. False Positive (FP): A legitimate URL incorrectly classified as phishing (false alarm). False Negative (FN): A phishing URL incorrectly classified as legitimate (missed detection).

Table IV. Confusion Matrix for Phishing URL Classification

	Predicted: Phishing	**Predicted: Legitimate**	**Total**
Actual: Phishing	True Positive (TP)	False Negative (FN)	TP + FN
Actual: Legitimate	False Positive (FP)	True Negative (TN)	FP + TN
Total	TP + FP	FN + TN	TP + TN + FP + FN

The following metrics are computed from the confusion matrix:

Precision (Eq. 1): The proportion of predicted phishing URLs that are genuinely malicious. High precision minimizes false alarms, reducing user disruption.

$$\textbf{Precision} = TP / (TP + FP) \quad \ldots (1)$$

Recall / Sensitivity (Eq. 2): The proportion of actual phishing URLs that are correctly identified. High recall minimizes missed detections, which is critical for security-oriented applications.

Recall = TP / (TP + FN) … (2)

F1-Score (Eq. 3): The harmonic mean of Precision and Recall, providing a single balanced metric that accounts for both false positives and false negatives. Particularly useful when class distribution is imbalanced.

F1-Score = 2 × (Precision × Recall) / (Precision + Recall) … (3)

Accuracy (Eq. 4): The overall fraction of URLs correctly classified across both classes. While intuitive, accuracy alone can be misleading on imbalanced datasets.

Accuracy = (TP + TN) / (TP + TN + FP + FN) … (4)

ROC-AUC (Eqs. 5–6): The Receiver Operating Characteristic (ROC) curve plots True Positive Rate (TPR, y-axis) against False Positive Rate (FPR, x-axis) across varying decision thresholds. The Area Under the Curve (AUC) summarizes classifier discrimination ability in a single scalar value between 0 and 1, where 1.0 represents perfect classification.

TPR = TP / (TP + FN) … (5) FPR = FP / (FP + TN) … (6)

Together, these metrics provide a comprehensive view of classifier performance, enabling fair comparison across studies that may use different dataset compositions or class distributions.

VII. OBSERVATIONS AND DISCUSSION

The following observations are drawn from a synthesis of the surveyed literature:

Dominance of Random Forest: Across the majority of reviewed studies, the Random Forest classifier consistently achieves the highest accuracy, often exceeding 97%, irrespective of dataset composition or split ratio. This superiority can be attributed to RF's ensemble nature, which mitigates overfitting through bagging and aggregates diverse decision tree predictions to produce robust outputs. The low false-negative rates associated with RF are particularly desirable in security-sensitive applications.

Impact of Training Data Volume: The findings of Singh and Siddavatam [12] demonstrate a clear monotonic relationship between training set size and classification accuracy. Models trained on 90% of available data outperform those trained on 50%, emphasizing the importance of large-scale, curated datasets for practical deployment.

Feature Selection as a Critical Preprocessing Step: Studies that incorporate dimensionality reduction—through techniques such as PCA, Information Gain, and RFE—report improved generalizability and reduced overfitting [15]. Eliminating redundant or correlated features reduces model complexity without sacrificing discriminative power.

Data Balancing Improves Reliability: Imbalanced datasets, where one class substantially outnumbers the other, can bias classifiers toward the majority class. Studies employing stratified or balanced sampling [13] demonstrate more reliable and interpretable evaluation metrics.

ANN as a Viable Alternative: While RF consistently leads in accuracy, Artificial Neural Networks (ANN) represent a competitive alternative, particularly when training time is not a constraint. ANN models exhibit strong performance in multi-dataset evaluations [14] and may offer advantages in capturing non-linear feature interactions.

Limitations of URL-Only Analysis: URL-based features alone may be insufficient against zero-day phishing attacks and adversarially crafted URLs designed to circumvent statistical detectors. Integrating content-based, visual, or behavioral features could enhance robustness.

Lack of Standardized Benchmarks: Variability in dataset sources, feature sets, and evaluation protocols across studies hinders direct comparison and reproducibility. The research community would benefit from the adoption of standardized benchmark suites.

VIII. CONCLUSION

This survey has systematically reviewed machine learning-based approaches to phishing URL detection, examining five representative studies in detail and synthesizing their findings across the dimensions of datasets, feature extraction, algorithm selection, and performance evaluation.

The evidence strongly supports the use of ensemble learning methods—particularly Random Forest—as the current state-of-the-art for phishing URL classification, offering consistent accuracy above 97% with favorable computational trade-offs. Preprocessing strategies including feature selection and class balancing are shown to significantly enhance model performance and generalizability.

Notwithstanding these advances, several research gaps remain. The field lacks universally adopted benchmark datasets, limiting reproducibility and cross-study comparisons. Adversarially resilient models that can withstand deliberate URL manipulation are underexplored. The integration of deep learning architectures—capable of learning hierarchical feature representations directly from raw URL strings—represents a promising frontier. Furthermore, the development of lightweight, real-time detection models suitable for resource-constrained environments warrants dedicated investigation.

It is the authors' intention that this survey serve as a structured reference for researchers and practitioners, facilitating the development of next-generation phishing detection systems that are accurate, adaptive, and operationally deployable.

REFERENCES

[1] Anti-Phishing Working Group (APWG), "Phishing Activity Trends Report, 4th Quarter 2020," 2021. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q4_2020.pdf
[2] Federal Bureau of Investigation, "Internet Crime Report 2020," Internet Crime Complaint Center (IC3), 2021. [Online]. Available: https://www.ic3.gov/Media/PDF/AnnualReport/2020_IC3Report.pdf
[3] Verizon, "2020 Data Breach Investigations Report," 2020. [Online]. Available: https://enterprise.verizon.com/resources/reports/2020-data-breach-investigations-report.pdf
[4] World Health Organization, "Cyber Security: Communicating for Health," WHO, Geneva, Switzerland. [Online]. Available: https://www.who.int/about/communications/cyber-security
[5] Y. Cao, W. Han, and Y. Le, "Anti-phishing based on automated individual white-list," in Proc. 4th ACM Workshop on Digital Identity Management (DIM '08), Alexandria, VA, USA, 2008, pp. 51–60.
[6] M. Sharifi and S. H. Siadati, "A phishing sites blacklist generator," in Proc. IEEE/ACS Int. Conf. on Computer Systems and Applications (AICCSA), Doha, Qatar, 2008, pp. 840–843.
[7] N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection based on associative classification data mining," Expert Systems with Applications, vol. 41, no. 13, pp. 5948–5959, Oct. 2014.
[8] L. Wenyin, G. Huang, L. Xiaoyue, Z. Min, and X. Deng, "Detection of phishing webpages based on visual similarity," in Special Interest Tracks and Posters of the 14th Int. Conf. on World Wide Web (WWW '05), Chiba, Japan, 2005, pp. 1060–1061.
[9] C. L. Tan, K. L. Chiew, et al., "Phishing website detection using URL-assisted brand name weighting system," in Proc. IEEE Int. Symp. on Intelligent Signal Processing and Communication Systems (ISPACS), Kuching, Malaysia, 2014, pp. 054–059.
[10] K. L. Chiew, E. H. Chang, and W. K. Tiong, "Utilisation of website logo for phishing detection," Computers & Security, vol. 54, pp. 16–26, Sep. 2015.
[11] K. M. Kumar and K. Alekhya, "Detecting phishing websites using fuzzy logic," Int. Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 5, no. 10, 2016.
[12] R. Mahajan and I. Siddavatam, "Phishing website detection using machine learning algorithms," Int. Journal of Computer Applications, vol. 181, no. 23, pp. 1–6, 2018.
[13] J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran, and B. S. Bindhumadhava, "Phishing website classification and detection using machine learning," in Proc. Int. Conf. on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020.
[14] M. Korkmaz, O. K. Sahingoz, and B. Diri, "Detection of phishing websites by using machine learning-based URL analysis," in Proc. 11th Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, Jul. 2020, pp. 1–7.
[15] M. N. Alam, D. Sarma, et al., "Phishing attacks detection using machine learning approach," in Proc. 3rd Int. Conf. on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 1–6.
[16] A. Subasi, E. Molah, F. Almkallawi, and T. J. Chaudhery, "Intelligent phishing website detection using Random Forest classifier," in Proc. Int. Conf. on Electrical and Computing Technologies and Applications (ICECTA), Ras Al Khaimah, UAE, Nov. 2017, pp. 1–5.
[17] Towards Data Science, "Phishing Domain Detection with ML," 2021. [Online]. Available: https://towardsdatascience.com/phishing-domain-detection-with-ml-5be9c99293e5
[18] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Phishing websites features," School of Computing and Engineering, Univ. of Huddersfield, UK, Tech. Rep., 2015.