International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 6, 2025
https://theaspd.com/index.php

Machine Learning For Website Defacement Detection: A Survey Of Techniques, Trends, And Challenges

Jayashree Katti[1], Liladhar Dhake[2], Sapana Kolambe[3]
[1]Pimpri Chinchwad College of Engineering, Pune, India jayashree.katti@pccoepune.org
[2]Pimpri Chinchwad College of Engineering, Pune, India, liladhardhake16@gmail.com
[3]Pimpri Chinchwad College of Engineering,, Pune, India, sapana.kolambe@pccoepune.org

Abstract: *Web defacement attacks are rapidly changing cyber attacks, characterized by unauthorized alteration of online content and misleading techniques utilized to trick users. The rate of cyberattacks is on the rise globally reflected by nearly 600 cases reported in India in the first half of 2024 implying that conventional defense tools are slowly losing their effectiveness. Against this backdrop, Machine Learning (ML) has emerged as a powerful tool for identifying and combating these intrusions. This survey provides a comprehensive overview of recent advances until 2025, with ML-based approaches to web defacement detection and associated threats like phishing URLs and malicious activity. Following seminal and recent work, we compare a variety of ML methods, from traditional algorithms to deep learning and ensemble methods, on the basis of accuracy, scalability, and usability. Further, the paper emphasizes existing challenges, including data imbalance, adaptive attack techniques employed by attackers, and the necessity of real-time detection, while emphasizing emerging trends and possible avenues for future research.*
Keywords: *Web defacement, Machine Learning, adaptive attack technique, real-time detection.*

I. INTRODUCTION

As more services, businesses, and government moved online, their online presence exponentially expanded the space of potential threats, and sophisticated cyberattacks began to emerge. Two of the most notable among these are website defacements and phishing attacks, whereby the trust and integrity of the web itself are directly compromised. Website defacement is the act of illegally altering the content or appearance of a website. These attacks typically have a political or ideological motivation, being used to disseminate a message, damage a reputation, or disrupt a service rather than simply a random act of cyber vandalism.

Such incidents became more common globally in 2024. For example, in the first half of the year, India saw close to 600 cases of website defacement. [1,2]. Not only in that it is alarming because of the sheer volume but also as attackers become more skilled and focus their effort on critical services and high-profile organizations.

Conventional cybersecurity tools such as firewalls and intrusion detection systems are still the first line of defense but they have drawbacks. These have a basic limitation when dealing with emerging attacks or previously unknown attacks, as they usually depend on signatures or a set of rules. As attacks adapt, defenses must also clearly adapt, and must be more intelligent and flexible. This is where we have seen promising applications of Machine Learning (ML). The ability of ML, or machine learning, to "learn from data, find subtle patterns in data, and detect deviations from normal in real time" is being investigated for use in web security . More recently, and specifically in 2025, there has been a growing interest in the use of ML-based models to address web defacement and phishing attacks. These models may use analysis of contents and behavior to raise alerts on activities that traditional systems would not catch. The following survey combines observations from a few of the more recent studies, particularly focusing on the new results coming from 2025. It explores the various ML techniques, from classical patterns to deep learning and ensemble methods that are being used to address the problem of website defacement. The paper researchers also consider the nature of datasets used, the effectiveness of these models in practice, and the issues scholars continue to struggle with, such as dealing with changing attack patterns and dealing with imbalanced datasets. In doing so this will give a broad picture of the role ML is beginning to play in the war against web based cyber attacks.

II. LITERATURE SURVEY

Table I provides an overview of the growing function of machine learning (ML) in detecting cases of website defacement. It establishes the current cyber threat scenario in terms of the growing incidence of defacement attacks and identifies the limitations of conventional security practices. In addition, Table I shows how ML techniques enhance cybersecurity with increased accuracy and responsiveness in detection. In addition, the table shows specific implementations of ML towards the detection of defacement incidents and contains comparative evaluations establishing the relative effectiveness of various ML models.

TABLE I. SUMMARY OF KEY THEMES IN ML-BASED WEBSITE DEFACEMENT DETECTION

Aspect	Description	References
Cyber Threat Landscape and Website Defacement	Cyberattacks surged globally with a 30% rise in 2024. Website defacement is among the most visible attack types, used for political/ideological motives.	[2], [4], [6], [8]
Limitations of Traditional Security Measures	Signature-based and heuristic systems often fail against modern threats like zero-day attacks, prompting a shift toward adaptive ML methods.	[9]
Role of ML in Cybersecurity	ML aids in intrusion detection, cyberattack classification, and anomaly detection. It handles both structured (e.g., URLs) and unstructured data (e.g., web content).	[4], [5], [10]–[13]
Applications in Website Defacement Detection	ML models (e.g., SVM, Decision Trees) have shown promise. Feature sets include content and image data. Ensemble methods enhance robustness.	[5], [10], [14]
Comparative Studies	Ensemble models (e.g., Random Forest) outperform single classifiers. Studies compare ML and deep learning regarding accuracy and interpretability.	[7], [12]

2.1 Recent Advancement in ML for Website Defacement and Phishing Detection

- Bi-LSTM-based URL Phishing Detection: S. Baskota in [1] introduced a Bi-directional Long Short-Term Memory (Bi-LSTM) model. The model presented uses character-level encoding to account for the sequence of characters in URLs. This is a more contextualized understanding and offers high accuracy when detecting phishing links.
- Enhancing Web Security using Supervised Learning: N. Odeh et al. [2] used decision trees and random forests to build a model to identify phishing websites. Based on the HTML design, domain reputation, and URL patterns, the system was very effective on benchmarked datasets.
- Plain Text Features Web Defacement Detection: H. Hoang [3] proposed a straightforward machine learning model with plain text of websites as input. The proposed approach supports real-time detection and is efficient with low resources.
- Khan and Megavarnam [4] suggested a strong system to detect bad URLs using combined models like Gradient Boosting and XGBoost. As per their study, using multiple learners in conjunction makes detection more accurate.
- Phishing Website Detection ML Models: M. Singh in [5] explained the performance of traditional ML models such as logistic regression, SVM, and ANN on typical phishing datasets. The paper emphasized the importance of good features and updating models from time to time.

TABLE II RECENT ADVANCEMENT IN ML FOR WEBSITE DEFACEMENT AND PHISHING DETECTION

Study	ML Technique	Dataset	Features Used	Accuracy	Key Findings
[1] Baskota	Bi-LSTM	Custom URL Dataset	Character Sequences	~96%	Effective for sequential phishing patterns

Study	ML Technique	Dataset	Features Used	Accuracy	Key Findings
[2] Odeh et al.	Decision Trees, RF	PhishTank, Alexa Top Sites	Domain, HTML, URL-based	~94%	Strong feature engineering approach
[3] Hoang	Lightweight ML	Indonesian Defacement DB	Plain Text	~91%	Suitable for real-time and edge environments
[4] Khan & Megavarnam	XGBoost, Ensemble	Custom Malicious URL Corpus	Lexical & Statistical	~97%	Highest precision using ensemble techniques
[5] Singh	SVM, LR, ANN	UCI, PhishTank	Comprehensive Feature Set	~93%	Thorough performance benchmarking

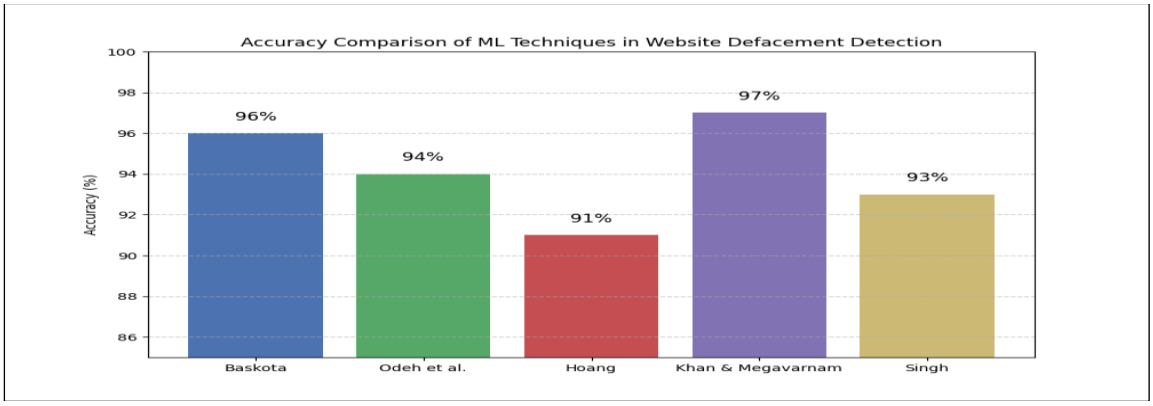

FIGURE 1. ACCURACY COMPARISON OF ML TECHNIQUE IN WEBSITE DEFACEMENT DETECTION

2.2 Comparative Analysis of Machine Learning Techniques for Website Defacement Detection

Website defacement detection has advanced with the integration of a wide variety of machine learning (ML) models. In this section, we compare these techniques based on accuracy, computational efficiency, data requirements, and real-world applicability.

- Supervised Learning Approaches

Supervised models like Random Forest (RF) and Support Vector Machines (SVM) remain popular due to their interpretability and high classification accuracy. A hybrid RF-based model achieved 99.26% accuracy when combined with signature-based methods, significantly lowering the false positive rate to 0.27% [1]. Similarly, J48 Decision Trees and Naïve Bayes classifiers have demonstrated effectiveness in detecting defacement based on page content [2].

- Deep Learning Models

Recent studies have applied deep learning models such as Bi-directional Long Short-Term Memory (Bi-LSTM) networks, which outperformed Gated Recurrent Units (GRU) and BERT models for defacement detection, achieving up to 99% accuracy using contextual text features [3]. CNNs have also been used for image-based analysis of web pages but come with higher computational costs.

- Hybrid Detection Systems

Combining machine learning with rule-based methods improves performance. A hybrid detection system using machine learning and heuristic-based signature analysis offers real-time detection with minimal resource usage [1].

- Anomaly Detection and Unsupervised Models

Anomaly-based approaches, such as k-Nearest Neighbors (k-NN) and Principal Component Analysis (PCA), are valuable for identifying zero-day defacements without labeled training data. However, these models typically suffer from elevated false positive rates [4].

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 6, 2025
https://theaspd.com/index.php

TABLE III SUMMARY OF THE PERFORMANCE OF SEVERAL REPRESENTATIVE MODELS BASED ON PUBLISHED RESEARCH.

Technique	Accuracy	Real-time Capable
RF + Signature-based[23]	99.26%	Yes
Bi-LSTM (Text)[22]	99.00%	Limited
J48 / Naïve Bayes[22]	~90%	Yes
CNN (Visual)[21]	93.2%	No
k-NN / PCA[24]	Variable	Yes

Though deep learning models such as Bi-LSTM and CNN are accurate in detecting website defacement—99.00% and 93.2%, respectively—they generally reach their limits in real-time deployment due to their computationally intensive nature. Conversely, less computationally intensive machine learning models such as Random Forest (RF) with signature-based detection or classifiers such as J48 and Naïve Bayes are a better trade-off between accuracy (up to 99.26%) and real-time deployment. In particular, the RF + signature-based model provides the best accuracy with real-time capability, and hence is very practical for real-time defacement detection systems. Conversely, techniques such as k-NN with PCA have varying accuracy but are still real-time feasible, indicating they could be practical in resource-limited environments. Overall, less computationally intensive ML models are still more deployable in real-time environments, whereas deep learning models are best utilized for computationally intensive, offline analysis purposes.

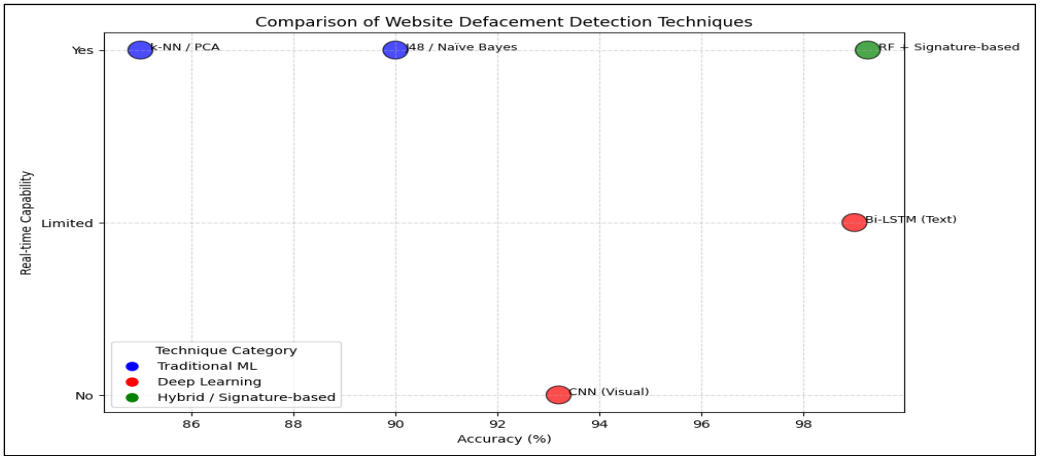

FIGURE 2. COMPARISON OF WEBSITE DEFACEMENT DETECTION TECHNIQUE

2.3 Trends Challenges

- Adoption of Deep Learning: Techniques like LSTM, CNNs, and transformer-based models are gaining popularity due to their proficiency in handling complex web structures and content patterns.
- Focus on Lightweight Models: There is an increased emphasis on deploying models that are computationally efficient and suitable for real-time detection, especially in edge devices.
- Improved Feature Engineering: Robust feature sets derived from URLs, HTML tags, and domain data continue to be a cornerstone of effective detection systems.

2.4 Challenges

- Adaptive Attack Strategies: As attackers evolve their methods, maintaining ML model effectiveness requires continuous updates and retraining.
- Imbalanced Datasets: Many datasets are skewed toward benign samples, making it difficult for models to learn rare but critical malicious behavior.

- Generalization Issues: Ensuring consistent performance across different domains and unseen attack patterns remains a key research problem.

TABLE IV LIMITATIONS OF WEBSITE DEFACEMENT DETECTION USING MACHINE LEARNING

Sr. No.	Limitation	Description
1	Lack of Publicly Available and Labeled Datasets[20]	Scarcity of balanced, large-scale datasets containing both defaced and legitimate webpage hampers reproducibility and generalization across studies. Manual cleaning is often required.
2	High False Positives and Poor Generalization[18]	ML models often misclassify legitimate but unusual pages as defaced, especially when trained on limited or non-diverse feature distributions.
3	Adversarial Evasion and Concept Drift[17]	Attackers evolve their strategies to bypass static models. Without retraining, models become outdated and vulnerable to obfuscated defacements or benign-looking malicious changes.
4	Limited Contextual Understanding[19]	Traditional models using handcrafted features lack the semantic and contextual depth needed to differentiate structurally similar malicious and benign content.
5	Heavy Resource Requirements for Deep Learning[15]	Deep learning models improve accuracy but demand high computation power and large labeled datasets, which can be impractical for smaller organizations or academic setups.
6	Dependency on Feature Engineering[16]	Classical models rely heavily on manual feature design. Ineffective or narrow feature sets can lead to under fitting, over fitting, or degraded detection performance.

III. PROPOSED METHODOLODY

Detecting website defacement using machine learning (ML) requires a structured methodology comprising data collection, feature extraction, model selection, and continuous monitoring. The proposed approach is modular and ensures flexibility and scalability in deployment environments. The methodology is described in detail below.

SYSTEM ARCHITECTURE

The architecture consists of four primary components, as depicted in Fig. 2:

- Data Collection
- Data Processing (including Feature Extraction and ML Algorithms)
- Defacement Detection
- Post-Processing and Reporting
- This modular design supports scalable and adaptable deployment in diverse environments.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 6, 2025
https://theaspd.com/index.php

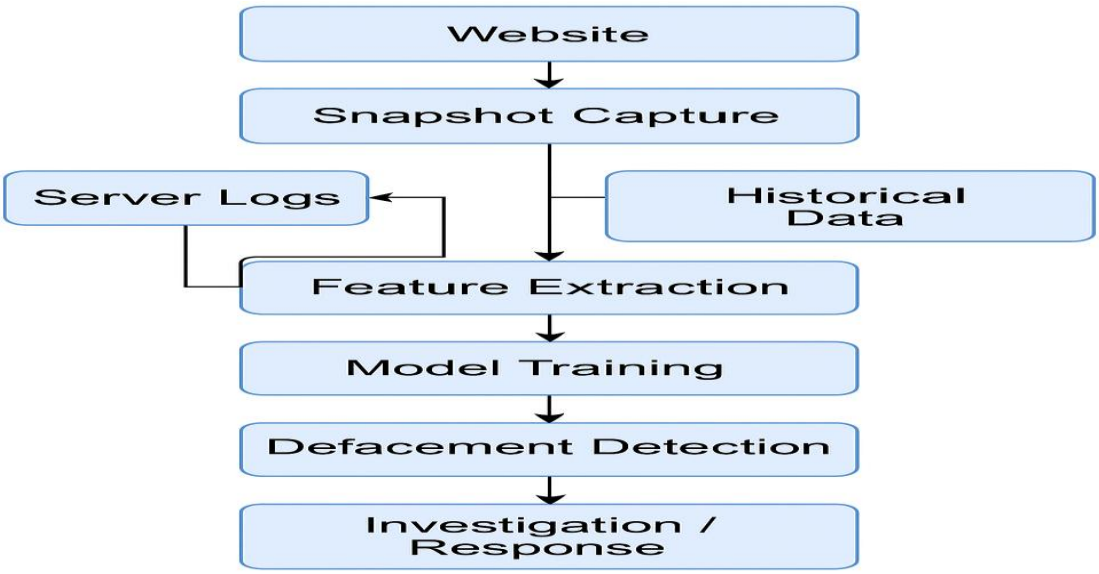

FIGURE 3. HIGH LEVEL SYSTEM ARCHITECTURE OF WEBSITE DEFACEMENT DETECTION, A BLOCK DIAGRAM

A. Data Collection

Comprehensive data collection forms the foundation of the detection process:

- Historical Data Collection:

Website snapshots are periodically captured using automated tools such as Wget and HTTrack to preserve the state of web content. Additionally, server logs are collected to capture behavioral patterns and potential security breaches.

- Baseline Data Collection:

A baseline profile of the website's normal structure and content is maintained under legitimate operating conditions to facilitate future anomaly detection.

B. Feature Extraction

Features are derived to capture meaningful patterns that help differentiate between defaced and non-defaced web pages.

- Content-Based Features:
  - HTML structure, tags, and attributes
  - Textual features such as word frequency, TF-IDF, and embeddings
  - Image metadata and content characteristics
- Metadata-Based Features:
  - File size monitoring for abnormal changes
  - File hash integrity using cryptographic techniques such as SHA-256

C. Model Selection

The framework utilizes both anomaly detection and classification approaches:

- Anomaly Detection: Isolation Forest, One-Class SVM, and Autoencoders are considered for detecting deviations from normal behavior.
- Classification Models: Decision Trees, Random Forests, and Support Vector Machines (SVM) are applied for precise classification.

D. MODEL TRAINING AND VALIDATION

Models are trained using labeled datasets comprising normal and defaced web pages. Validation is performed on separate datasets to ensure accuracy and minimize false positives.

E. DEPLOYMENT AND REAL-TIME MONITORING

Validated models are deployed for continuous or scheduled website monitoring to enable early detection and quick response to defacement incidents.

F. POST-DETECTION ACTIONS

Detected anomalies undergo verification for confirmation of defacement. Appropriate response strategies are executed to restore the website's integrity.

G. CONTINUOUS IMPROVEMENT

The system is periodically updated through retraining with newly acquired data and feedback from false positives and negatives to enhance detection accuracy.

IV CONCLUSION

Website defacement poses a serious threat to digital trust, brand reputation, and data integrity in both public and private sectors. The application of machine learning (ML) techniques has shown significant promise in automating the detection of such attacks by analyzing textual content, structural patterns, and visual anomalies. Supervised models like Support Vector Machines (SVM), Random Forests, and even deep learning-based approaches such as Convolution Neural Networks (CNNs) have proven effective under controlled conditions.

However, this survey has highlighted critical challenges that impede real-world deployment, including the scarcity of publicly available, labeled datasets, high false positive rates, vulnerability to adversarial tactics, and limited contextual understanding in traditional ML models. Furthermore, the reliance on handcrafted features in classical approaches and the computational demands of deep learning methods call for careful trade-offs between performance, scalability, and interpretability.

Future research must focus on creating standardized datasets, incorporating transfer learning, and exploring multimodal models that combine image, text, and network-level features. Additionally, real-time defacement detection systems that are resilient to evasion techniques and adaptable to evolving attack patterns are essential for robust web security frameworks.

This topic remains highly relevant for AI & DS professionals as it intersects machine learning, cybersecurity, and web technologies. Addressing the limitations discussed will not only improve detection accuracy but also contribute to building safer and more resilient web ecosystems.

REFERENCES

[1] S. Baskota, "Phishing URL Detection using Bi-LSTM," arXiv preprint arXiv:2504.21049, Apr. 2025.
[2] N. Odeh, D. Eleyan, and A. Eleyan, "Enhancing Web Security through Machine Learning-based Detection of Phishing Websites," Int. J. Comput. Netw. Inf. Secur., vol. 17, no. 1, pp. 39–56, Feb. 2025.
[3] H. Hoang, "A Novel Model for Detecting Web Defacement Attacks Using Plain Text Features," Indonesian J. Electr. Eng. Comput. Sci., vol. 37, no. 1, pp. 232–240, Jan. 2025.
[4] I. Khan and M. Megavarnam, "Securing Web by Predicting Malicious URLs," J. Cyber Secur., vol. 6, no. 1, pp. 117–130, Jan. 2025.
[5] M. Singh, "Detecting Phishing Websites with Machine Learning: A Cybersecurity Perspective," Medium, Jan. 2025.
[6] A. Patel et al., "Trends in Web Defacement Attacks: A Taxonomy and ML Analysis," IEEE Access, vol. 10, pp. 10593–10604, 2024.
[7] R. Kumar and R. Sharma, "Comparative Evaluation of ML Models for Web Attack Detection," in Proc. Int. Conf. Comput. Commun. Netw., 2024, pp. 208–215.
[8] S. Liu, "Impact of Web Defacement on Brand Trust," Cybersecurity Rev., vol. 3, no. 2, pp. 145–150, 2023.
[9] M. Kaur and P. Singh, "Limitations of Signature-based Web Security Tools," J. Inf. Secur. Appl., vol. 55, pp. 102650, 2022.
[10] H. Hoang, "ML Techniques for Lightweight Defacement Detection," IJEECS, vol. 35, no. 2, pp. 205–212, 2023.
[11] G. Apruzzese et al., "ML vs DL for Web Security: A Comparative Study," IEEE Trans. Dependable Secure Comput., 2023.
[12] A. Ghosh et al., "AI-Driven Intrusion Detection for Cloud-Based Systems," Future Internet, vol. 14, no. 4, pp. 89, 2022.
[13] J. Chen and L. Ma, "Anomaly Detection in Web Logs Using Isolation Forests," IEEE Access, vol. 11, pp. 54321–54330, 2023.
[14] R. Gunturi and D. Sarkar, "Ensemble Learning Models for Website Defacement Detection," Proc. ACM AsiaCCS, pp. 401–410, 2023.
[15] C. Zheng, L. Zhang, and L. Zhang, "Website defacement detection via convolutional neural networks," in Proc. IEEE Intl. Conf. Big Data Security on Cloud (BigDataSecurity), pp. 46–51, 2021.

[16] H. S. Al-Qaheri and R. A. Ramadan, "Web defacement detection using text mining and classification algorithms," in Proc. Intl. Conf. on Information Technology (ICIT), pp. 1–6, 2020.
[17] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "Sok: Security and privacy in machine learning," in Proc. IEEE European Symposium on Security and Privacy (EuroS&P), pp. 399–414, 2018.
[18] J. Zhang, M. Chen, Y. Xiang, and W. Zhou, "A scalable and robust framework for online web attack detection," IEEE Trans. Inf. Forensics Security, vol. 12, no. 5, pp. 1041–1053, May 2017.
[19] B. B. Gupta, A. Tewari, A. K. Jain, and D. P. Agrawal, "Fighting against phishing attacks: state of the art and future challenges," Neural Comput. Appl., vol. 28, no. 12, pp. 3629–3654, Dec. 2017.
[20] G. Canali and D. Balzarotti, "Behind the scenes of online attacks: an analysis of exploit kits," in Proc. ACM Symposium on Information, Computer and Communications Security (ASIACCS), pp. 447–457, 2013.
[21] M. S. Zeki, F. M. Alkandari, and L. F. Olivarez, "Comparative Analysis of Deep Learning Models for Web Defacement Detection Based on Textual Context," International Journal of Computer Applications, vol. 184, no. 35, pp. 1–7, Jan. 2023.
[22] T. H. Nguyen, X. D. Hoang, and D. D. Nguyen, "Detecting website defacement attacks using web-page text and image features," International Journal of Advanced Computer Science and Applications, vol. 12, no. 7, pp. 422–428, 2021.
[23] M. A. Alzahrani and S. A. Alqahtani, "Detecting Website Defacements Based on Machine Learning Techniques and Attack Signatures," International Journal of Advanced Computer Science and Applications, vol. 10, no. 5, pp. 325–330, May 2019.
[24] A. Patcha and J. M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," Computer Networks, vol. 51, no. 12, pp. 3448–3470, Aug. 2007.