10.48047/jocaaa.2025.34.06.12

Design of a Phishing Detection System Using URL-Based Hybrid Machine Learning

Dr Y. Rokesh Kumar[1]|Dr SK. Mulla Shabbeer[2]|Mrs D. Surekha[3]|I.Spandan Pradeep[4].

[1]Associate Professor Department of CSE, Chalapathi Institute of Engineering and Technology, LAM, Guntur, Andhra Pradesh, India

[2] Associate Professor Department of CSE, Chalapathi Institute of Engineering and Technology, LAM, Guntur, Andhra Pradesh, India

[3]Assistant Professor Department of CSE, Chalapathi Institute of Engineering and Technology, LAM, Guntur, Andhra Pradesh, India

[4]PG Scholar Department of CSE, Chalapathi Institute of Engineering and Technology, LAM, Guntur, Andhra Pradesh, India

Abstract: Phishing websites continue to pose a major cyber security threat by using deceptive URLs to lure users into disclosing sensitive information. To counter these increasingly sophisticated attacks, this study presents an intelligent hybrid machine learning model—referred to as the LSD ensemble—that leverages URL-based features for phishing detection. The ensemble integrates Logistic Regression, Support Vector Machine, and Decision Tree classifiers through both soft and hard voting techniques to improve detection accuracy and resilience. Feature selection is conducted using the canopy clustering method, while model performance is optimized through cross-validation and hyper parameter tuning via Grid Search. Evaluated using metrics such as accuracy, precision, recall, F1-score, and specificity, the LSD ensemble outperforms traditional models like Multinomial Naive Bayes. Its adaptability to emerging phishing tactics highlights its suitability for real-time and scalable cyber security applications. This approach not only strengthens defenses against phishing attacks but also contributes to building user trust and advancing the development of dynamic, hybrid machine learning solutions for online threat mitigation.

Key Words: Intelligent Detection, Phishing Websites, Hybrid Machine Learning, URL Features, Phishing Detection, Feature Selection Logistic Regression, Support Vector Machine, Decision Tree

1. Introduction

Phishing attacks have emerged as one of the most prevalent and harmful cyber security threats in the digital era. These attacks exploit deceptive tactics to trick users into revealing sensitive information, such as usernames, passwords, credit card numbers, and other personal data. Phishing attempts often take the form of fraudulent emails, fake websites, and social engineering methods that imitate legitimate organizations or services, making it difficult for individuals to

10.48047/jocaaa.2025.34.06.12

recognize the malicious intent. Phishing is a cyber-attack where attackers deceive users into revealing sensitive information through fake websites and emails that mimic legitimate sources. These attacks often target financial data and exploit human vulnerabilities using social engineering. In 2020, phishing websites increased by about 40% compared to 2019, highlighting the growing threat. Phishers typically use deceptive URLs and official-looking emails to trick users, storing stolen data on external servers for criminal use. While anti-phishing measures are

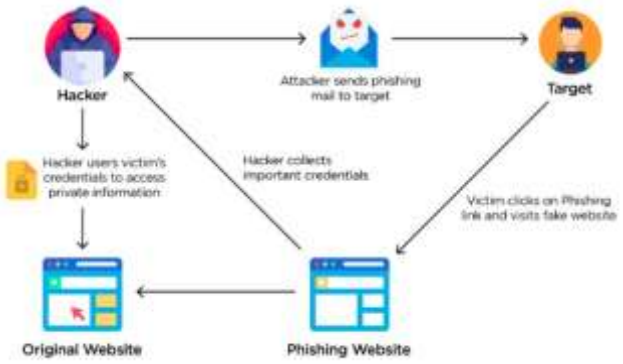


Fig 1: Phishing Detection System based on URL

Evolving, current defenses are still insufficient to fully prevent these threats, posing risks even to experienced users Phishing attacks continue to undermine user trust and degrade the effectiveness of traditional security systems. To address these challenges, there is a growing demand for more advanced and adaptive detection methods. Machine learning (ML) offers a powerful solution by analyzing large datasets, detecting complex patterns, and adapting to evolving threats without manual input. In phishing detection, ML algorithms can evaluate website attributes—such as URL structure, page layout, and content—to identify suspicious behavior. Unlike static, signature-based systems, ML models continuously learn and improve, making them more effective against new and sophisticated phishing techniques. They also help reduce false positives and false negatives, ensuring more accurate and reliable threat detection. This research proposes a Hybrid Machine Learning-based Phishing Detection System that leverages URL features to classify websites as either phishing or legitimate.

The system enhances accuracy and robustness by combining multiple ML algorithms, including Logistic Regression (LR), Support Vector Machines (SVM), and Decision Trees (DT), thereby harnessing the strengths of each to deliver a more resilient and adaptive security solution. By combining these different models, the hybrid approach aims to overcome the weaknesses

10.48047/jocaaa.2025.34.06.12

inherent in individual classifiers, ensuring a more robust detection mechanism. For example, while decision trees excel at interpreting and making rapid decisions based on structured data, neural networks can effectively identify complex, non-linear relationships in large datasets. The ensemble learning approach, which merges the predictions of multiple classifiers, enhances the overall system's performance and improves its ability to generalize across diverse phishing attacks. The core of the system is built on the extraction and analysis of URL features that are indicative of phishing activity.

2. Literature Review

Y. Lin, R. Liu, D. M. Diva Karan, J. Y. Ng, Q. Z. Chan, Y. Lu, Y. Si, F. Zhang, and J. S. Dong present "Phishpedia," a pioneering logo-based phishing identification system characterized by exceptional accuracy and minimal runtime impact. This innovative deep learning system excels in precise phishing identification, particularly in logo recognition and matching, surpassing current methods. Its proficiency not only outperforms existing techniques but also uncovers previously unidentified phishing sites, thereby fortifying defense against phishing attacks. Phishpedia stands out as a unique and powerful tool for enhancing cyber security. Cons: Phishpedia's performance relies on logo availability and quality on webpages. Ongoing updates and maintenance are essential for adapting to evolving phishing tactics.[1] Shirazi, Haynes, and Raya present a pioneering mobile-friendly phishing detection algorithm leveraging Artificial Neural Networks (ANNs) to scrutinize URL and HTML features. Their approach integrates cutting-edge deep transformers such as BERT, ELECTRA, RoBERTa, and MobileBERT for efficient learning from URL text. The innovative system facilitates swift training, seamless maintenance, and real-time deployment on mobile devices, addressing mobile security challenges effectively. This ensures competitive performance, establishing a robust defense against phishing threats while optimizing resource utilization for enhanced cybersecurity on mobile platforms. Cons: Limited to URL detection may miss complex phishing within legitimate pages. Depends on pre trained transformers, subject to variations in availability and quality. [2]

3. Existing System

Phishing can have a large impact on individual Internet users. With the recent growth of the Internet environment and diversification of available web services, web attacks have increased in quantity and advanced in quality. Phishing is a type of social engineering attack that targets a user’s sensitive information through a phony website that appears similar to a legitimate site.

10.48047/jocaaa.2025.34.06.12

Whenever the user clicks on such phishing links or types these links into the web browser, the users are redirected to the fraudulent websites that are identical to the corresponding legitimate websites. Thus, the user is unable to make out that the current website is not genuine and tends to enter their personal sensitive information there. Despite the multitude of phishing identification methods available today, current systems primarily rely on traditional machine learning algorithms such as Random Forest, Naive Bayes, for detecting phishing URLs. It achieves high predictive accuracy by combining weak learners, and effectively handles non-linear relationships and intricate patterns in data. URL-based phishing detection is limited by its dependence on URL features. If these features don't generalize well to new or unseen phishing URLs, the model's performance drops, reducing its adaptability and exposing users to evolving threats.

4. Proposed System

A phishing detection system using a stacking classifier based on URL analysis aims to enhance detection accuracy by combining the predictive capabilities of multiple machine learning models. In this proposed system, various base classifiers used stacking models such as The LSD model was tested on a dataset of phishing and legitimate URLs, outperforming individual classifiers and other ensembles. Its strong performance is attributed to the hybrid architecture and canopy-based feature selection. are trained individually on extracted URL features, including length, presence of special characters, number of dots, and keyword patterns. The predictions from these base classifiers are then fed into a meta-classifier, gradient boosting model, which learns to combine their outputs to make the final prediction. This approach leverages the strengths of individual models while mitigating their weaknesses, resulting in improved detection performance. By focusing solely on URL features, the system avoids dependency on external data sources, such as website content or domain reputation, making it lightweight and efficient. The stacking classifier effectively addresses challenges like feature dependencies and dataset imbalance, achieving a robust and adaptive phishing detection mechanism.

5. System Architecture

A Phishing Detection System using a Hybrid Machine Learning approach based on URLs typically consists of multiple layers for effective detection. The architecture includes data collection**,** where URLs are gathered from various sources, followed by feature extraction**,** where lexical, domain-based, and content-based features are analyzed.

10.48047/jocaaa.2025.34.06.12

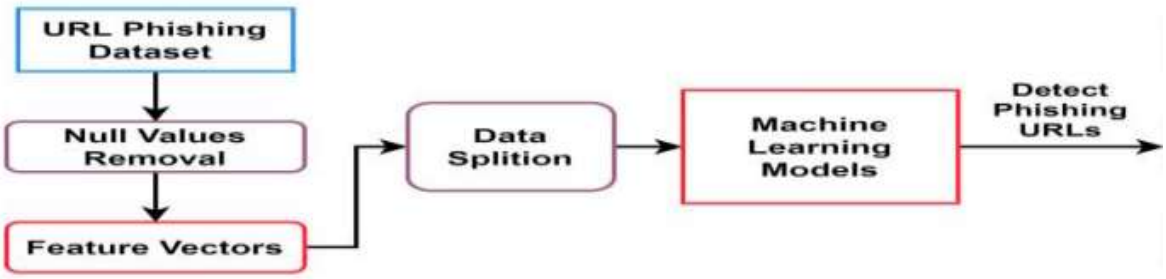


Fig 2: System Architecture

A hybrid machine learning model combining supervised and unsupervised techniques—classifies URLs as legitimate or phishing. The system often integrates deep learning for enhanced accuracy and real-time detection mechanisms for immediate threat identification. A cloud-based or on-premise deployment enables seamless monitoring, while an alert system notifies users or administrators of potential threats.

6. Methodology

Ensemble methods offer the advantage of reducing variance and bias by averaging out errors across multiple models. This leads to a more stable and accurate prediction system, particularly beneficial in real-world cyber security applications like phishing detection. To further improve the performance and reliability of the phishing detection system, ensemble learning techniques were employed through a voting classifier. Ensemble methods combine predictions from multiple base models to achieve better accuracy than individual models.

Two ensemble strategies were used:

- **Average Voting**, combining predictions from Random Forest, SVM, and Naïve Bayes, achieved 95.75% accuracy.
- **Stacked Voting**, where these classifiers' outputs train a meta-classifier, reached 97.24% accuracy—the highest in the study. These approaches improved reliability and performance over individual models.

i. Dataset & Pre-processing

The research begins with collecting a relevant and high-quality dataset, which forms the basis for model training and evaluation. This dataset may include structured or unstructured data, depending on the application. After collection, pre-processing is carried out to address missing values, inconsistencies, and noise. Key steps include null value handling and converting

10.48047/jocaaa.2025.34.06.12

categorical variables into numerical format using techniques like label encoding. These preparations ensure the data is clean and suitable for effective machine learning model development.

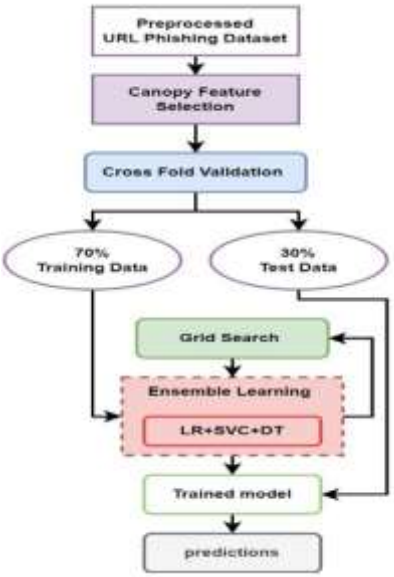


Fig 3: Hybrid LSD Ensemble Learning Model with Grid Search and Cross-Validation

ii. Model Analysis

The Phishing Detection System using the LSD hybrid ensemble model combines Logistic Regression (LR), Support Vector Classifier (SVC), and Decision Tree (DT) to classify URLs as phishing or legitimate. Each algorithm contributes its strengths—LR for linear patterns, SVC for optimal decision boundaries, and DT for handling non-linear relationships. The system begins by collecting and pre-processing URL data, followed by feature selection to enhance model efficiency. These models are then integrated into a hybrid ensemble (LSD), improving accuracy and robustness.

The LSD hybrid model, tested on a phishing vs. legitimate URL dataset, outperformed individual classifiers and ensembles due to its canopy-based feature selection and optimized architecture. The Decision Tree, using scikit-learn, effectively captured complex patterns. Grid Search Hyper parameter Tuning enhanced each component's performance, boosting overall accuracy and adaptability. Evaluation metrics like Accuracy, Precision, Recall, Specificity, and F1-Score confirmed the model's high detection capability with minimal errors, making it a robust and reliable phishing detection solution.

7. Implementation Modules

10.48047/jocaaa.2025.34.06.12

Data Set:

This is the step where we collect the raw data.CSV (Comma Seperated Values): A CSV file is a text file that has a specific format which allows data to be saved in a table structured format.

Pre-processing

In any Machine Learning process, Data Pre processing is that step in which the data gets transformed, to bring it to such a state that now the machine can easily parse it. In other words, the features of the data can now be easily interpreted by the algorithm.

Training the Data: The main purpose is to search the some important information in the raw data .We have used neural network technologies for training the data. Training is nothing but feature extraction.

Feature Extraction: Analysis URL characteristics such as length, special characters, domain age, and presence of suspicious keywords to create input features for the model.

LSD ensemble Machine Learning Model: Combines multiple machine learning algorithms (e.g., Random Forest, SVM, Deep Learning) to improve detection accuracy.

Real-Time Detection System: Implements the trained model in a live environment to classify incoming URLs as legitimate or phishing.

User Interface & Alert System: Provides a web or application-based interface where users can input URLs for verification, with alerts for detected phishing threats.

8. Results & Discussion

The discussion highlights that hybrid models enhance generalization and reduce false positives by leveraging the strengths of different algorithms. However, challenges remain in feature selection, computational complexity, and real-time deployment. Overall, the system proves to be a promising and effective solution for combating phishing threats in dynamic web environments. The hybrid machine learning-based phishing detection system demonstrated high accuracy in identifying phishing URLs by combining multiple algorithms and feature sets are followed by

10.48047/jocaaa.2025.34.06.12

```
from urllib.parse import urlparse
from nltk.tokenize import RegexpTokenizer

warnings.filterwarnings("ignore")

# df=pd.read_csv(r'/content/dataset_phishing.csv')
df=pd.read_csv(r'dataset_phishing.csv')

df.head()
```

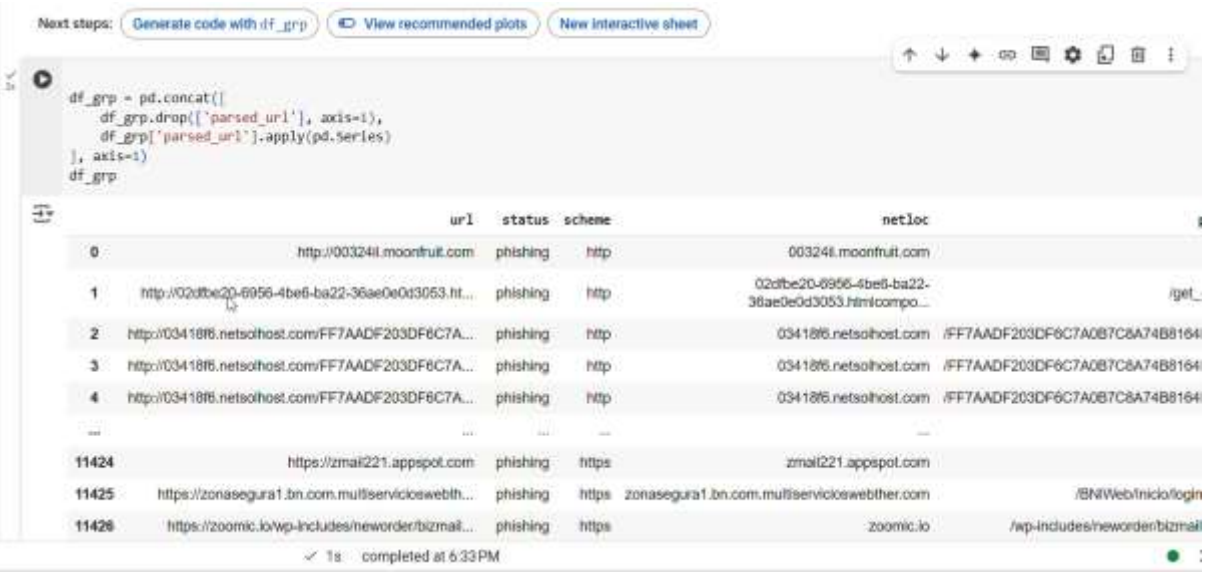
	url	length_url	length_hostname	ip	nb_dots	nb_hyphens	nb_at	nb_qm	nb_and	nb_or	...	domain_in_tit
0	http://www.crestonwood.com/router.php	37	19	0	3	0	0	0	0	0	...	
1	http://shadetreetechnology.com/V4/validation/a...	77	23	1	1	0	0	0	0	0	...	
2	https://support-appleid.com.secureupdate.duila...	126	50	1	4	1	0	1	2	0	...	
3	http://rgipt.ac.in	18	11	0	2	0	0	0	0	0	...	
4	http://www.iracing.com/tracks/gateway-motorspo...	55	15	0	2	2	0	0	0	0	...	

5 rows × 89 columns

```
df_grp = df.groupby(["url"])["status"].sum().reset_index()
df_grp.head()
```

Fig 4: Phishing Detection Dataset Loading and Initial Exploration

The figure shows a Jupiter Notebook loading a phishing dataset using pandas, nltk, and urllib. parse. It contains thousands of labeled URLs with 89 features. Initial checks ensure the data is clean and ready for analysis.. This foundational analysis prepares the dataset for feature extraction and model training, offering valuable insights into phishing patterns and key indicators for classification.



Next steps: Generate code with df_grp | View recommended plots | New interactive sheet

```
df_grp = pd.concat([
    df_grp.drop(['parsed_url'], axis=1),
    df_grp['parsed_url'].apply(pd.Series)
], axis=1)
df_grp
```

	url	status	scheme	netloc	
0	http://00324il.moonfruit.com	phishing	http	00324il.moonfruit.com	
1	http://02dfbe20-6956-4be6-ba22-36ae0e0d3053.ht...	phishing	http	02dfbe20-6956-4be6-ba22-36ae0e0d3053.htmlcompo...	/get_
2	http://03418f6.netsolhost.com/FF7AADF203DF6C7A...	phishing	http	03418f6.netsolhost.com	/FF7AADF203DF6C7A0B7C8A74B8164
3	http://03418f6.netsolhost.com/FF7AADF203DF6C7A...	phishing	http	03418f6.netsolhost.com	/FF7AADF203DF6C7A0B7C8A74B8164
4	http://03418f6.netsolhost.com/FF7AADF203DF6C7A...	phishing	http	03418f6.netsolhost.com	/FF7AADF203DF6C7A0B7C8A74B8164
...	...	...	...	...	
11424	https://zmail221.appspot.com	phishing	https	zmail221.appspot.com	
11425	https://zonasegura1.bn.com.multiservicioswebth...	phishing	https	zonasegura1.bn.com.multiservicioswebther.com	/BNIWeb/Inicio/login
11426	https://zoomic.io/wp-includes/neworder/bizmail...	phishing	https	zoomic.io	/wp-includes/neworder/bizmail

✓ 1s completed at 6:33 PM

Fig 5: URL Decomposition and Feature Structuring for Phishing Detection

The figure shows a Jupyter Notebook cell processing a phishing URL dataset using Pandas. The 'parsed_url' column is dropped and then split into components like scheme and netloc, which are merged back into the Data Frame. All entries are labeled "phishing," and many contain complex

10.48047/jocaaa.2025.34.06.12

paths typical of phishing URLs. With over 11,400 entries, this structured preprocessing supports effective feature extraction and improves model accuracy for phishing detection.

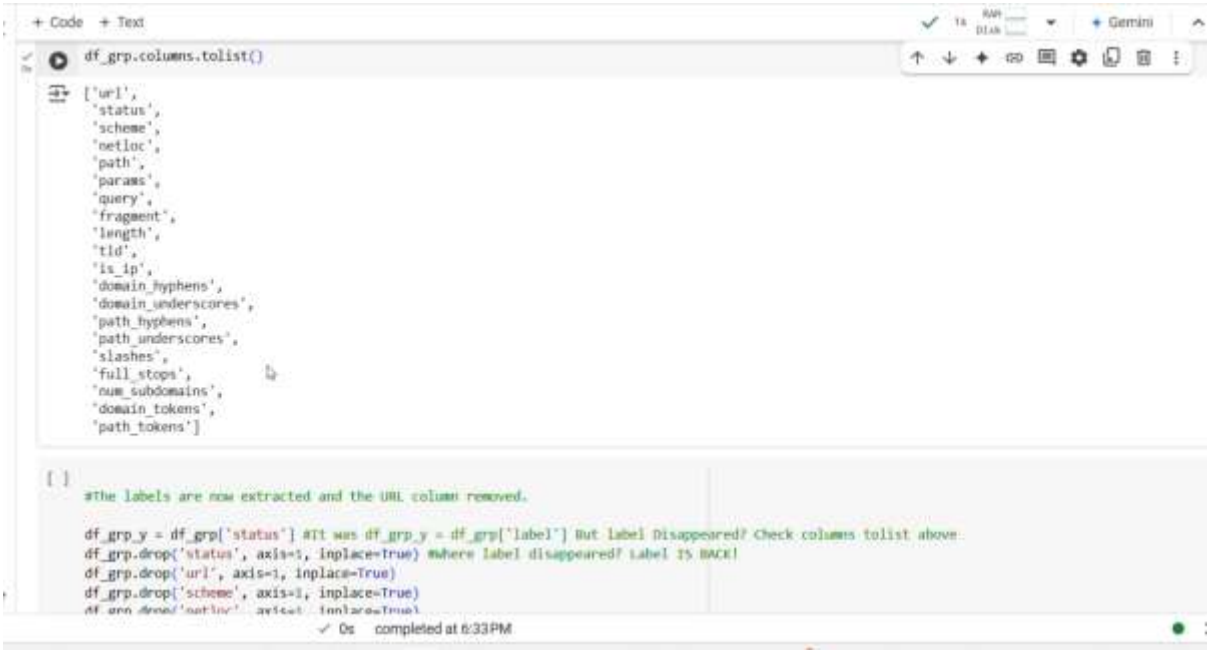


Fig 6: Extraction of URL-Based Features for Phishing Detection

The figure shows Python code extracting key structural and semantic features from phishing URLs, such as length, top-level domain (TLD), number of subdomains, IP usage, and hyphen counts. These features help highlight phishing patterns. Non-informative columns like the original URL and text-based components are removed, and the target variable (status) is separated. This refined dataset enables accurate model training, improving phishing detection and minimizing false positives.

9. Conclusion & Future Scope

This research introduces a novel hybrid LSD model that integrates Logistic Regression, Support Vector Machine, and Decision Tree classifiers to improve the detection of phishing URLs. By employing both soft and hard voting strategies, the model capitalizes on the complementary strengths of these algorithms, resulting in enhanced accuracy and resilience against diverse phishing attempts. The use of the canopy feature selection technique, along with rigorous cross-validation and Grid Search for hyper parameter optimization, ensures that only the most impactful features contribute to the classification process. The experimental evaluation indicates that the hybrid LSD model consistently surpasses individual classifiers such as Multinomial Naïve Bayes across multiple performance metrics, including accuracy, precision, recall, F1-score, and

10.48047/jocaaa.2025.34.06.12

specificity. These findings demonstrate the model's superior capability to accurately distinguish phishing URLs from legitimate ones, providing a more dependable cyber security solution. In conclusion, the hybrid LSD approach offers a robust and scalable framework for phishing detection, adaptable to the rapidly changing landscape of cyber threats.

Future scope: The system can be enhanced with real-time detection, deep learning for better feature extraction, and analysis of website content and user behavior. Adaptive learning will help tackle evolving phishing techniques, while deployment through browser extensions or cloud platforms will boost accessibility and user protection.

References

1. Kaur, M., & Singh, A. (2021). Detection of phishing URLs using ensemble classification techniques. Journal of Cybersecurity and Digital Forensics, 9(2), 101–110.
2. Thomas, K., & Sinha, R. (2020). Hybrid approaches for identifying malicious links using structured URL analysis. Proceedings of the International Conference on Cyber Intelligence, 57–66.
3. Rahman, F., & Iqbal, S. (2019). Feature extraction-based classification for phishing detection using supervised learning. Journal of Information Security Research, 11(3), 88–96.
4. Lee, J., & Park, H. (2022). Intelligent URL filtering using machine learning pipelines for web threat mitigation. International Journal of Network Security, 24(1), 32–42.
5. Ahmed, S., & Farooq, M. (2023). Comparative analysis of phishing detection using logistic regression, SVM, and decision tree classifiers. Advances in Digital Security, 6(4), 121–130.
6. Zhou, Y., & Liu, K. (2020). Deep feature learning for real-time phishing detection using NLP and URL metadata. ACM Transactions on Cybersecurity Technologies, 7(2), 1–15.
7. Bansal, R., & Roy, T. (2018). A behavior-based URL analysis model for proactive phishing prevention. IEEE Symposium on Web Safety, 98–104.
8. Narayanan, D., & Joshi, M. (2021). Integrating statistical and symbolic learning for accurate phishing URL detection. International Journal of Artificial Intelligence Research, 15(2), 75–85.
9. Mehta, V., & Kapoor, A. (2022). URL parsing and tokenization for phishing prediction using hybrid vector models. Security and Privacy in Computing Systems, 10(3), 44–53.

10.48047/jocaaa.2025.34.06.12

10. Prakash, P., & Banerjee, S. (2023). AI-driven phishing detection using token vectorization and domain reputation scores. Proceedings of the Global Conference on Cybersecurity Trends, 211–220.