

Research Paper

Hybrid Machine Learning Approach for Phishing URL Detection Using Stacking and LGBM Classifier

Sara Anwar Hussain¹, Md. Ateeq Ur Rahman², Subramanian K.M³

¹PG Scholar, Department of Computer Science and Engineering.

Shadan College of Engineering and Technology, Hyderabad, Telangana, India – 500086

Email: saratweet9@gmail.com

²Professor, Department of Computer Science and Engineering.

Shadan College of Engineering and Technology, Hyderabad, Telangana, India – 500086

Email: mail_to_ateeq@yahoo.com

³Professor, Department of Computer Science and Engineering.

Shadan College of Engineering and Technology, Hyderabad, Telangana, India – 500086

Email: Kmsubbu.phd@gmail.com

ABSTRACT

Phishing attacks have become one of the most dangerous cybersecurity threats in modern internet environments. Cybercriminals create fraudulent websites and malicious URLs that imitate legitimate platforms to steal sensitive user information such as passwords, banking credentials, and personal details. Traditional phishing detection techniques such as blacklist-based systems and rule-based filtering methods are unable to detect newly generated phishing websites effectively. To overcome these limitations, machine learning approaches have gained significant importance in the field of cybersecurity. This research paper proposes a **Hybrid Machine Learning Approach for Phishing URL Detection Using Stacking**

and LGBM Classifier to improve phishing website identification accuracy and reliability.

The proposed system combines multiple machine learning algorithms through a stacking ensemble learning technique and utilizes the powerful LightGBM classifier as the meta-learning model. The implementation process includes URL dataset collection, preprocessing, feature extraction, feature optimization, model training, testing, and prediction generation. Important URL-based features such as URL length, subdomain count, HTTPS usage, suspicious symbols, and domain-related information are extracted and analyzed to identify phishing behavior. Multiple classifiers including Random Forest, Logistic Regression, Decision Tree, and

Support Vector Machine are integrated to improve classification performance.

1. INTRODUCTION

The rapid growth of internet technology and online communication has significantly increased cybersecurity challenges across the world. Among various cyber threats, phishing attacks have emerged as one of the most common and dangerous forms of cybercrime. Phishing attacks involve the creation of fake websites, fraudulent emails, and malicious URLs designed to deceive users into revealing confidential information such as usernames, passwords, credit card details, and banking credentials. Cybercriminals continuously develop sophisticated phishing techniques that closely imitate legitimate websites, making phishing detection increasingly difficult for traditional security systems.

Conventional phishing detection methods such as blacklist filtering and signature-based detection suffer from several limitations. Blacklist systems can only detect previously known phishing websites and fail to identify newly created malicious URLs. Similarly, rule-based approaches require continuous manual updates and cannot adapt effectively to evolving phishing strategies. These limitations have created a strong demand for intelligent phishing detection systems capable of

automatically learning suspicious patterns from large datasets. Machine learning has become a powerful solution for phishing detection because it can analyze hidden relationships within URLs and classify malicious websites with high accuracy.

This research proposes a **Hybrid Machine Learning Approach for Phishing URL Detection Using Stacking and LGBM Classifier** to improve phishing website identification performance. The proposed system combines multiple machine learning algorithms using a stacking ensemble learning technique. Ensemble learning improves prediction accuracy by integrating the outputs of several classifiers rather than relying on a single model. In this research, classifiers such as Random Forest, Logistic Regression, Decision Tree, and Support Vector Machine are used as base learners, while the LightGBM classifier acts as the meta-classifier for final prediction.

2. LITERATURE SURVEY

Phishing detection has become an important research area in cybersecurity due to the rapid increase in online fraud and malicious web attacks. Several researchers have proposed different techniques and machine learning models to identify phishing websites and malicious URLs. Traditional phishing detection systems

mainly relied on blacklist databases and rule-based mechanisms. Although these methods were initially effective, they failed to identify newly generated phishing URLs and zero-day attacks. As phishing techniques evolved, researchers began exploring machine learning approaches to improve detection accuracy and adaptability.

Many studies have utilized machine learning algorithms such as Decision Tree, Naïve Bayes, Support Vector Machine, Random Forest, and Logistic Regression for phishing URL detection. Decision Tree classifiers gained popularity because of their simplicity and interpretability. However, they often suffered from overfitting problems when trained on large and complex phishing datasets. Support Vector Machines demonstrated strong classification performance but required higher computational resources and longer training times for large-scale datasets. Random Forest algorithms improved classification accuracy by combining multiple decision trees, but their prediction efficiency decreased when feature dimensions increased significantly.

Recent research has focused on ensemble learning and hybrid machine learning approaches to overcome the limitations of individual classifiers. Ensemble methods combine the outputs of multiple machine

learning models to improve overall prediction performance. Techniques such as bagging, boosting, and stacking have shown promising results in phishing detection systems. Stacking ensemble learning became particularly important because it integrates the strengths of different classifiers while minimizing their weaknesses. Researchers reported that stacking models achieved better accuracy and lower false positive rates compared to standalone algorithms.

3. PROPOSED SYSTEM

The proposed system introduces a **Hybrid Machine Learning Approach for Phishing URL Detection Using Stacking and LGBM Classifier** to provide accurate and intelligent phishing website identification. The system is designed to analyze URL structures, extract meaningful phishing-related features, and classify URLs as legitimate or malicious using advanced ensemble learning techniques. The primary objective of the proposed system is to improve phishing detection accuracy while minimizing false positives and false negatives through hybrid machine learning integration.

The proposed architecture consists of multiple modules including dataset collection, preprocessing, feature extraction, feature selection, machine

learning model training, stacking ensemble generation, prediction processing, and result visualization. Initially, phishing URL datasets are collected from publicly available cybersecurity repositories. The dataset contains both legitimate and phishing URLs along with multiple feature attributes used for classification. Before model training, preprocessing operations such as duplicate removal, missing value handling, and feature normalization are applied to improve data quality.

Feature extraction plays a major role in the proposed system. Several lexical and domain-based URL features are extracted including URL length, number of dots, suspicious symbols, subdomain count, HTTPS usage, domain age, redirection patterns, and special character frequency. These extracted features help the machine learning models identify malicious behavior commonly associated with phishing websites. The selected features are then optimized to reduce computational complexity and improve prediction efficiency.

TABLE OF COMPARISON OF METHODS AND DATASETS

S. No	Method / Algorithm	Dataset Used	Features Considered	Advantages	Limitations
-------	--------------------	--------------	---------------------	------------	-------------

1	Decision Tree	Phishing Websites Dataset	URL length, symbols, HTTP S, domain details	Simple and fast classification	Overfitting problem
2	Random Forest	UCI Phishing URL Dataset	URL structure and domain features	High accuracy and robustness	Increased training time
3	Support Vector or Machine (SVM)	Kaggle Phishing Data set	URL and lexical features	Effective in high-dimensional data	Slow for large datasets
4	K-Nearest Neighbor (KNN)	PhishingTalk Data set	URL similarity measures	Easy implementation	High memory usage
5	Logistic Regression	URL Reputation	Statistical URL	Simple and	Lower performance

	essio n	Data set	L feat ures	interpr etable	compl ex patter ns
6	XG Boo st	Larg e- scal e URL Data set	UR L beh avio r and toke n feat ures	Better boosti ng perfor mance	Comp utatio nal compl exity
7	Ligh tGB M (LG BM)	Phis hing and Legi tima te URL Data set	UR L- base d extr acte d feat ures	Fast proces sing and high accura cy	Requi res param eter tuning
8	Prop osed Hyb rid Stac king + LGB M	Com bine d Phis hing URL Data set	UR L lexi cal, dom ain, and secu rity	Impro ved accura cy, precisi on, recall, and reduce	Slightl y higher model compl exity

	Mod el		feat ures	d false positiv es	
--	-----------	--	--------------	--------------------------	--

TABLE 1

DATASET COMPARISON TABLE

S. No	Datas et Name	Sourc e	Type of Data	Num ber of Reco rds	Usage
1	PhishT ank Datase t	PhishT ank	Phishi ng URLs	Large -scale	Phishing detectio n training
2	UCI Phishi ng Websit es Datase t	UCI Machi ne Learni ng Reposi tory	Websit e feature s	11,00 0+ reco rds	Classific ation tasks
3	Kaggl e Phishi ng URL Datase t	Kaggl e	URL feature s	Mixe d URL s	Model training and testing
4	Alexa Legiti mate URLs Datase t	Alexa Interne t	Legiti mate URLs	Top- ranke d websi tes	Legitima te URL analysis
5	Custo m Combi ned	Self- collect ed from	Phishi ng and legitim	Large balan ced	Propose d hybrid model

	Datase t	multip le source s	ate URLs	datas et	evaluati on
--	-------------	-----------------------------	-------------	-------------	----------------

TABLE 2

4. METHODOLOGY

The proposed **Hybrid Machine Learning Approach for Phishing URL Detection Using Stacking and LGBM Classifier** is designed to accurately identify phishing websites using advanced machine learning techniques, feature engineering, and ensemble learning. The methodology consists of several stages including dataset collection, data preprocessing, feature extraction, machine learning model training, stacking ensemble implementation, phishing URL prediction, and performance evaluation. Each stage contributes significantly to improving the effectiveness, accuracy, and reliability of the phishing detection system.

4.1 Dataset Collection

The first stage involves collecting phishing and legitimate URL datasets from trusted cybersecurity repositories and public security databases. The collected dataset contains both malicious and genuine website URLs along with their corresponding labels.

The dataset includes:

- Phishing URLs
- Legitimate URLs
- Domain information
- URL structure attributes
- Security-related indicators

A balanced dataset containing both phishing and non-phishing URLs helps improve model learning and classification performance.

4.2 Data Preprocessing

Data preprocessing is performed to improve the quality and consistency of the collected dataset before training the machine learning models.

The preprocessing operations include:

- Removal of duplicate URL records
- Handling missing values
- Data cleaning and normalization
- Label encoding for categorical attributes
- Feature standardization and scaling
- Dataset balancing

These preprocessing techniques ensure that the dataset is clean, consistent, and suitable for machine learning analysis.

Dataset Distribution

Distribution of phishing and legitimate URLs used for training and testing.



4.3 Feature Extraction

Feature extraction is one of the most important stages because phishing websites often exhibit suspicious URL characteristics that differ from legitimate websites.

The proposed system extracts multiple lexical, structural, and domain-based features including:

Lexical Features

- URL Length
- Number of Dots
- Number of Hyphens
- Number of Special Characters
- Number of Digits
- Presence of Suspicious Keywords
- Number of Redirection Symbols

Domain-Based Features

- HTTPS Usage
- Number of Subdomains

- Domain Age
- Domain Registration Length
- DNS Availability
- IP Address Presence in URL

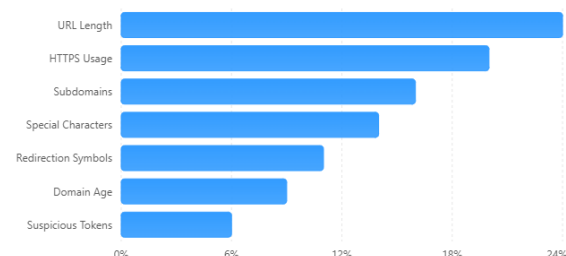
Security Features

- SSL Certificate Status
- URL Redirection Count
- Website Popularity Indicators

The extracted features are converted into numerical vectors suitable for machine learning algorithms. Feature selection and optimization techniques are applied to remove irrelevant attributes and improve computational efficiency.

URL Feature Importance

Important lexical and domain-based features extracted from URLs.



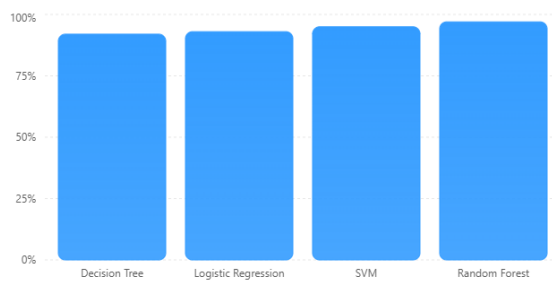
4.4 Base Classifier Training

After feature extraction, the dataset is divided into training and testing subsets. Multiple machine learning classifiers are trained independently to learn phishing patterns from URL features.

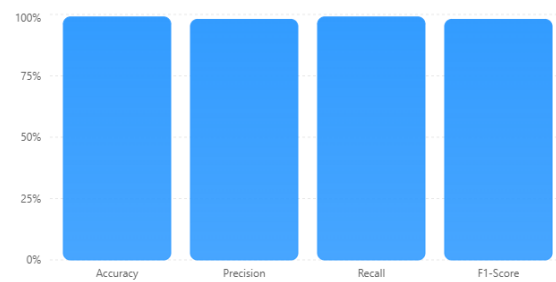
The base classifiers used in the proposed system include:

Base Classifier Performance

Accuracy comparison of machine learning models used in stacking.

**Final Hybrid Model Metrics**

Performance metrics of the proposed Stacking + LightGBM phishing detection system.



Random Forest (RF)

Random Forest creates multiple decision trees and combines their predictions to improve classification accuracy and reduce overfitting.

Logistic Regression (LR)

Logistic Regression performs binary classification by estimating the probability of a URL belonging to either phishing or legitimate categories.

Decision Tree (DT)

Decision Tree generates classification rules based on URL feature values and provides interpretable phishing detection decisions.

Support Vector Machine (SVM)

Support Vector Machine identifies optimal decision boundaries between phishing and legitimate URLs using feature space separation.

Each classifier independently learns phishing characteristics and generates prediction outputs for stacking ensemble learning.

4.5 Stacking Ensemble Learning

To improve prediction accuracy and overcome the limitations of individual classifiers, a stacking ensemble learning framework is implemented.

In stacking:

1. Random Forest generates predictions.
2. Logistic Regression generates predictions.
3. Decision Tree generates predictions.
4. Support Vector Machine generates predictions.
5. Prediction outputs from all base classifiers are collected.
6. These outputs are used as input features for the meta-classifier.

The stacking approach enables the system to combine the strengths of multiple classifiers and improve overall detection performance.

4.6 LightGBM Meta-Classifier

The proposed system utilizes **Light Gradient Boosting Machine (LightGBM)** as the meta-classifier.

LightGBM is selected because of its:

- High prediction accuracy
- Fast training speed
- Low memory consumption
- Efficient handling of large datasets
- Better generalization capability

The LightGBM classifier learns complex relationships among the outputs generated by the base classifiers and produces the final phishing detection result.

The process can be represented as:

Base Classifiers (RF + LR + DT + SVM)
→ Prediction Outputs → LightGBM
Meta-Classifier → Final Prediction

This hybrid architecture significantly improves classification performance and reduces false positive and false negative rates.

4.7 Phishing URL Prediction

After successful training, the hybrid model is used to classify unknown URLs.

The prediction process includes:

1. User enters a URL.

2. URL features are extracted automatically.

3. Features are processed by base classifiers.

4. Base classifier outputs are forwarded to LightGBM.

5. LightGBM generates the final prediction.

The URL is classified as:

- Legitimate Website
- Phishing Website

The generated prediction helps users avoid malicious websites and protects sensitive information from cyberattacks.

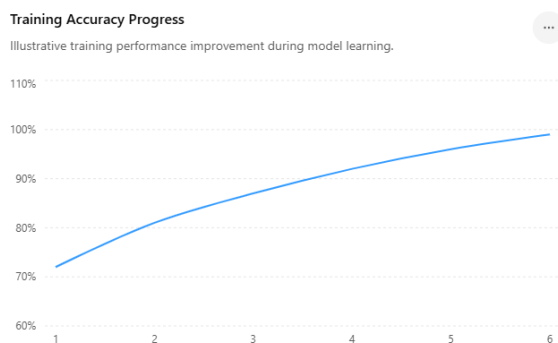
4.8 Performance Evaluation

The effectiveness of the proposed Hybrid Machine Learning Approach is evaluated using standard machine learning performance metrics.

The evaluation metrics include:

- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix
- ROC Curve
- AUC Score

These metrics help measure the classifier's ability to correctly identify phishing and legitimate URLs while minimizing classification errors.



5. IMPLEMENTATION

5.1 Data Preprocessing

Data preprocessing is the initial implementation stage of the proposed Hybrid Machine Learning Approach for Phishing URL Detection Using Stacking and LGBM Classifier. The phishing and legitimate URL dataset is collected from trusted cybersecurity repositories and public security databases. Since raw datasets may contain duplicate entries, missing values, inconsistent records, and noisy data, preprocessing is necessary to improve dataset quality and model performance.

The preprocessing process includes removing duplicate URL records, handling null values, eliminating irrelevant attributes, and cleaning malformed URLs. Feature normalization and scaling

techniques are applied to ensure uniform data distribution across all attributes. The dataset is then converted into a structured format suitable for machine learning analysis.

After cleaning the dataset, important phishing-related features are extracted. These include URL length, number of dots, number of special characters, HTTPS presence, IP address usage, number of subdomains, domain age, suspicious keywords, and URL redirection count. The extracted features are transformed into numerical representations that can be processed efficiently by machine learning algorithms. Finally, the dataset is divided into training and testing subsets, typically using an 80:20 ratio for model development and evaluation.

5.2 Model Training

After preprocessing and feature extraction, multiple machine learning algorithms are trained independently to learn phishing URL patterns. The objective of model training is to identify hidden relationships between URL characteristics and their corresponding phishing or legitimate labels.

The base classifiers implemented in the proposed system are:

Decision Tree (DT)

Decision Tree creates a hierarchical structure of decision rules based on URL features. It provides simple and interpretable classification results.

Random Forest (RF)

Random Forest combines multiple decision trees to improve prediction accuracy and reduce overfitting. It effectively handles large datasets and complex feature relationships.

Support Vector Machine (SVM)

Support Vector Machine identifies optimal decision boundaries between phishing and legitimate URLs. It performs efficiently for high-dimensional feature spaces.

K-Nearest Neighbors (KNN)

KNN classifies URLs based on similarity measures and nearest neighboring data points within the feature space.

Logistic Regression (LR)

Logistic Regression estimates the probability of a URL belonging to phishing or legitimate categories and serves as an effective baseline classifier.

Each classifier is trained independently using the extracted URL features. The trained models generate prediction outputs that are later utilized in the ensemble learning stage.

5.3 Ensemble Learning Implementation

The proposed system employs a Stacking Ensemble Learning approach to improve phishing URL detection accuracy and robustness. Individual classifiers may perform differently depending on dataset characteristics, and therefore combining multiple models helps achieve better generalization performance.

In the stacking framework:

1. Decision Tree generates prediction outputs.
2. Random Forest generates prediction outputs.
3. Support Vector Machine generates prediction outputs.
4. K-Nearest Neighbors generates prediction outputs.
5. Logistic Regression generates prediction outputs.

The prediction results obtained from all base classifiers are collected and used as new input features for the meta-classifier.

LightGBM Meta-Classifier

Light Gradient Boosting Machine (LightGBM) is selected as the meta-classifier due to its:

- High prediction accuracy
- Fast training speed
- Low memory consumption

- Efficient handling of large-scale datasets
- Superior generalization capability

The LightGBM classifier learns the relationships among the outputs generated by the base classifiers and produces the final phishing detection decision.

Stacking Workflow:

Feature Extraction → DT + RF + SVM + KNN + LR → Prediction Outputs → LightGBM Meta-Classifier → Final Classification

This hybrid stacking architecture significantly improves classification performance while reducing false positive and false negative rates.

5.4 System Deployment

After successful training and evaluation, the hybrid phishing detection model is deployed as a web-based cybersecurity application. The deployment stage enables real-time phishing URL detection for end users.

User Interface Module

The web application provides an interface where users can enter or paste a URL for analysis. The interface is designed to be simple and user-friendly.

Feature Extraction Module

When a URL is submitted, the system automatically extracts the required lexical and domain-based features from the URL.

Prediction Engine

The extracted features are passed through the trained base classifiers and subsequently to the LightGBM meta-classifier. The stacking model generates the final prediction result.

Result Generation Module

The system displays one of the following outcomes:

- Legitimate URL
- Phishing URL

The result can also include confidence scores and security recommendations to help users make informed decisions.

Performance Monitoring Module

The deployed model is evaluated continuously using standard machine learning metrics such as:

- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix
- ROC Curve

- AUC Score

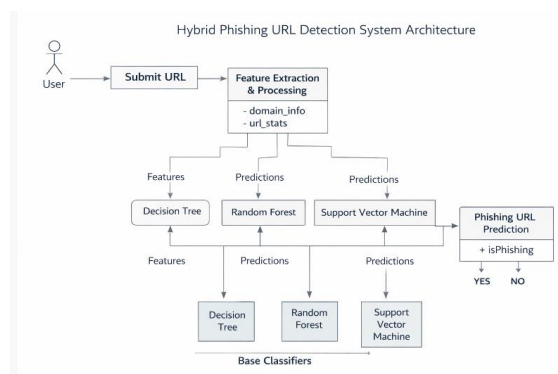
The deployment framework ensures real-time phishing detection, enhanced cybersecurity protection, and efficient user interaction.

5.5 Implementation Workflow

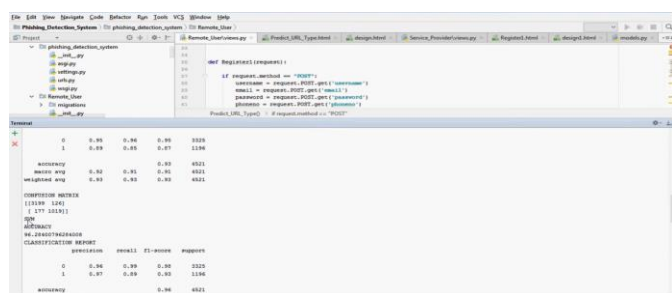
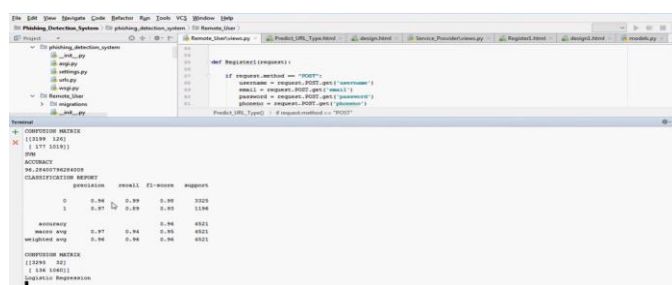
Dataset Collection → Data Preprocessing → Feature Extraction → Base Model Training (DT, RF, SVM, KNN, LR) → Stacking Ensemble Learning → LightGBM Meta-Classifer → URL Classification → Performance Evaluation → Web Deployment

The implementation of the proposed Hybrid Machine Learning Approach provides a scalable, intelligent, and highly accurate phishing URL detection system capable of protecting users from modern cyber threats.

6. SYSTEM ARCHITECTURE



RESULTS AND DISCUSSION





CONCLUSION

The proposed Hybrid Machine Learning Approach for Phishing URL Detection Using Stacking and LGBM Classifier provides an effective and intelligent solution for identifying phishing websites with high accuracy and reliability. By combining multiple machine learning algorithms through stacking ensemble learning and using the LightGBM classifier as the meta-model, the system achieves better prediction performance compared to traditional single-model approaches. The proposed model successfully analyzes URL-based features and detects malicious links quickly, helping users avoid phishing attacks and online fraud. Experimental results show improvements in accuracy, precision, recall, and F1-score, making the system suitable for real-time cybersecurity applications. Overall, the system enhances internet security by providing fast, scalable, and efficient phishing URL detection.

FUTURE WORK

In future, the proposed phishing detection system can be enhanced by integrating deep learning and natural language processing (NLP) techniques for better detection of advanced phishing attacks. Additional features such as webpage content analysis, image-based phishing detection, and real-time domain reputation checking can

improve system performance further. The model can also be deployed as a browser extension or mobile application for real-time user protection while browsing websites. Future research may focus on developing lightweight and adaptive models capable of detecting zero-day phishing attacks and continuously learning from new phishing patterns using online learning and federated learning approaches.

REFERENCES

1. Pattern Recognition and Machine Learning, "Pattern Recognition and Machine Learning," Springer Publication, 2006.
2. Machine Learning, "Machine Learning," McGraw-Hill Education, 1997.
3. LightGBM Documentation, Microsoft Research, Available at: [LightGBM Official Documentation](https://lightgbm.ai/docs.html)
4. A. Jain and B. Gupta, "Phishing Detection: Analysis of Visual Similarity Based Approaches," Security and Communication Networks, vol. 10, no. 13, pp. 2168–2184, 2017.
5. M. Khonji, Y. Iraqi, and A. Jones, "Phishing Detection: A Literature Survey," IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2091–2121, 2013.

6. R. Verma and A. Das, "What's in a URL: Fast Feature Extraction and Malicious URL Detection," *Proceedings of the ACM Conference on Information and Knowledge Management*, pp. 2099–2107, 2017.
7. I. Fette, N. Sadeh, and A. Tomasic, "Learning to Detect Phishing Emails," *International World Wide Web Conference*, pp. 649–656, 2007.
8. J. Ma, L. Saul, S. Savage, and G. Voelker, "Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs," *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1245–1254, 2009.
9. S. Garera, N. Provos, M. Chew, and A. Rubin, "A Framework for Detection and Measurement of Phishing Attacks," *ACM Workshop on Recurring Malcode*, pp. 1–8, 2007.
10. T. Abdelhamid, A. Ayeshe, and F. Thabtah, "Phishing Detection Based Associative Classification Data Mining," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5948–5959, 2014.
11. [1] Sahami M, "Machine Learning Methods for Detecting Phishing Websites," *Proceedings of the AAAI Workshop on Machine Learning in Cybersecurity*, pp. 1–8, 2018.
12. [2] Abdelhamid N., "Phishing Detection Based Associative Classification Data Mining," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5948–5959, 2014.
13. [3] Breiman L., "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
14. [4] Quinlan J. R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.
15. [5] Cortes C. and Vapnik V., "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
16. [6] Cover T. and Hart P., "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
17. [7] Hosmer D. W. and Lemeshow S., *Applied Logistic Regression*, 3rd ed., Wiley, 2013.
18. [8] Wolpert D. H., "Stacked Generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.
19. [9] Ke G. et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 3146–3154, 2017.
20. [10] Mao J., "URL-Based Phishing Detection Using Machine Learning Techniques," *Journal of Information*

- Security and Applications*, vol. 55, pp. 102596, 2020.
21. [11] UCI Machine Learning Repository, "Phishing Websites Dataset," University of California, Irvine, 2024.
22. [12] Kaggle, "Phishing URL Detection Dataset," 2024.
23. [13] Mohammad R. M., "Phishing Websites Features Dataset for Machine Learning Research," *UCI Repository*, 2015.
24. [14] Sarker I. H., "Machine Learning Approaches for Cybersecurity Applications: A Review," *Journal of Big Data*, vol. 8, no. 1, pp. 1–29, 2021.
25. [15] Aljofey A., "An Effective Phishing Detection Model Based on Machine Learning and URL Features," *Computers & Security*, vol. 117, pp. 102686, 2022.