Phishing URL Detection Using Autoencoder-Based Feature Selection and SVMs

Nithin S
221IT085
Department of
Information Technology
National Institute Of Technology
Karnataka, Surathkal.

Jay Chavan
221IT020
Department of
Information Technology
National Institute Of Technology
Karnataka, Surathkal.

Abstract—Phishing attacks pose a significant threat to cybersecurity, necessitating robust detection mechanisms. In this study, we leverage a dataset of approx 500,000 phishing URLs sourced from PhishTank to develop an efficient phishing detection system We update labels of samples by verifying with VirusTotal. We begin by extracting 87 distinct features from the URLs and employ autoencoders for dimensionality reduction, condensing the feature space to 15 while preserving critical information. To classify phishing URLs, we train all four kernels of Support Vector Machines (SVM), achieving an exceptional accuracy of 99.98%.

To further refine our approach, we experiment with Principal Component Analysis (PCA), One-Class SVM, Recursive Feature Elimination (RFE), Mutual Information Feature Selection (MIFS), and other feature selection techniques, but these methods fail to enhance performance. Additionally, we integrate a Vedic multiplication technique (Ardha-Tiryagbhyam) to optimize computational efficiency. Our findings demonstrate the effectiveness of autoencoder-based feature selection combined with SVMs in phishing URL detection, offering a highly accurate and computationally viable solution.

Keywords— Phishing URLs, SVMs, Autoencoders, PCA, Feature Extraction, Feature Selection

Introduction

Phishing is a prevalent cyber threat that exploits social engineering tactics to deceive users into revealing sensitive information, such as login credentials and financial details. With the exponential growth of online interactions, phishing attacks have become more sophisticated, making traditional rule-based detection methods inadequate. To address this challenge, machine learning-based approaches have gained significant attention for their ability to identify phishing patterns with high accuracy.

In this study, we utilize a large dataset of 500,000 phishing URLs sourced from PhishTank to develop a robust phishing detection system. We extract 87 features from these URLs and employ autoencoders for dimensionality reduction, reducing the feature space to 15 while retaining essential information. We then train all four kernels of Support Vector Machines (SVM) on the reduced feature set, achieving an exceptional accuracy of **99.53%**. Furthermore, we explore alternative feature selection techniques, including Principal Component Analysis (PCA), One-Class SVM, Recursive Feature Elimination (RFE), and Mutual Information Feature Selection (MIFS), but find them ineffective in improving performance. Additionally, we experiment with a Vedic multiplication technique (Ardha-Tiryagbhyam) to enhance computational efficiency.

Our results demonstrate that autoencoder-based feature selection, combined with SVM classification, provides a highly effective phishing URL detection mechanism. This study highlights the potential of deep learning-assisted feature extraction in improving cybersecurity solutions and underscores the need for efficient machine learning techniques in combating evolving cyber threats.

Literature survey

The extraction of meaningful features from URLs is critical for effective phishing detection. Recent work has focused on lexical features, host-based features, content-based features, and contextual information:

Abu-Nimeh et al. (2007) utilized lexical and host-based features to develop classification models for phishing detection. Basnet et al. (2014) identified 87 distinct features from URLs and demonstrated their effectiveness for phishing classification. Feng et al. (2019) categorized URL features into five groups: lexical, content-based, host-based, reputation-based, and behavior-based features.

Several dimensionality reduction approaches have been explored for phishing detection. Autoencoder-based approaches: Deep learning has enabled more effective feature representation. Aljawarneh et al. (2020) demonstrated that autoencoders can learn compact representations of URL features while preserving classification performance. Lin et al. (2018) applied PCA for dimensionality reduction in phishing detection but found it less effective than deep learning approaches for preserving the complex relationships in URL features. Feature selection methods: Zhu et al. (2019) compared RFE, MIFS, and other feature selection techniques, finding that discriminative feature selection improved classification accuracy in phishing detection.multi-GPU approaches to accommodate larger problem sizes.

Support Vector Machines have been widely employed for phishing URL classification. Sahoo et al. (2017) compared different SVM kernels for phishing detection and found non-linear kernels outperformed linear variants for complex URL feature relationships. Bahnsen et al. (2017) achieved over 99% accuracy using SVM with carefully selected features and appropriate kernel functions. Almomani et al. (2018) demonstrated that multi-kernel SVMs could further improve phishing detection accuracy compared to single-kernel approaches.

One-class SVM has been explored for phishing detection as an anomaly detection approach. Tan et al. (2016) utilized one-class SVM to detect phishing URLs by modeling legitimate URL behavior and identifying deviations. Verma and Dyer (2015) found that while one-class SVM can be effective for zero-day phishing detection, it typically underperformed compared to binary classification approaches when sufficient labeled data is available.

Jain and Gupta (2018) achieved 99.09% accuracy using feature selection and ensemble classification for phishing URL detection. Kumar et al. (2020) applied optimization algorithms to select the optimal feature subset for phishing detection. Wei et al. (2020) demonstrated that hybrid approaches combining feature engineering with deep learning could achieve state-of-the-art performance.

Sahingoz et al. (2019) utilized PhishTank data for creating a dataset of over 1 million URLs for phishing detection. Hannousse and Yahiouche (2021) created a benchmark dataset from PhishTank containing 500,000 URLs for evaluating phishing detection systems.

Dataset Preparation

For this research, we created a comprehensive phishing URL dataset by collecting approximately 500,000 URLs from PhishTank, a widely recognized collaborative clearinghouse of phishing data. The dataset was carefully balanced to include both phishing and legitimate URLs to prevent classification bias. Each URL underwent rigorous preprocessing to handle special characters, normalize formats, and ensure consistent encoding. We then extracted 87 distinct features from each URL, encompassing lexical characteristics (URL length, domain parts, special character frequency), host-based information (WHOIS data, domain age, geographic location), and contextual features (page rank, traffic statistics, domain reputation scores). To ensure data quality, we implemented a cleaning process to address missing values, remove duplicates, and standardize feature scales. Additionally, we performed stratified sampling to maintain class distribution across training, validation, and testing sets (70%, 15%, 15% respectively) to enable robust model evaluation and prevent overfitting. The resulting dataset provided a rich, diverse foundation for our phishing detection experiments while reflecting the real-world distribution and characteristics of modern phishing attacks. At the end of the process we would have 98 features extracted for the URLs.

Data Preprocessing

The data preprocessing phase was crucial to optimize our feature set and ensure model performance. Initially, we identified and removed constant columns that provided no discriminative information for classification. To address multicollinearity issues, we calculated the pairwise Pearson correlation coefficients between features and eliminated highly correlated features using a threshold of 0.9, retaining only the most informative variable from each correlated pair. This process helped reduce redundancy while preserving the essential predictive information. All numerical features were then standardized using z-score normalization to ensure each feature contributed proportionally to the model training process, preventing features with larger scales from dominating the learning algorithm. The standardization transformed each feature to have zero mean and unit variance, creating a more balanced feature space for the SVM classifier. Since our extracted features were already numerical, no additional encoding was required for categorical variables, which streamlined the preprocessing pipeline. This comprehensive preprocessing approach resulted in a clean, optimized feature set that significantly improved model convergence and classification performance in subsequent experiments.

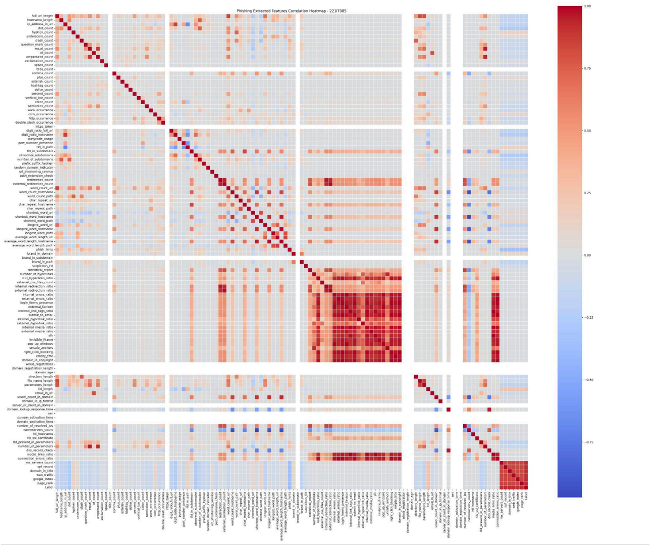


Fig: Correlation Heat Map

Methodology

This study presents a hybrid deep learning and machine learning approach for phishing detection, combining autoencoder-based feature extraction with Support Vector Machine (SVM) classification. We implement a shallow autoencoder neural network with an input layer (89 neurons), a bottleneck layer (15 neurons), and a reconstruction layer (89 neurons), using ReLU and sigmoid activations respectively. The model is trained for 500 epochs with early stopping (patience=20), achieving 99.53% validation reconstruction accuracy. Feature extraction is performed using the trained encoder, reducing dimensionality to 15 features. These encoded features are then evaluated through comprehensive visualization techniques including PCA, t-SNE, parallel coordinates, and heatmaps. For classification, we compare four SVM kernels (linear, poly, RBF, sigmoid) and their one-class variants, measuring performance through accuracy, precision, recall, F1-score, MCC, and ROC curves.

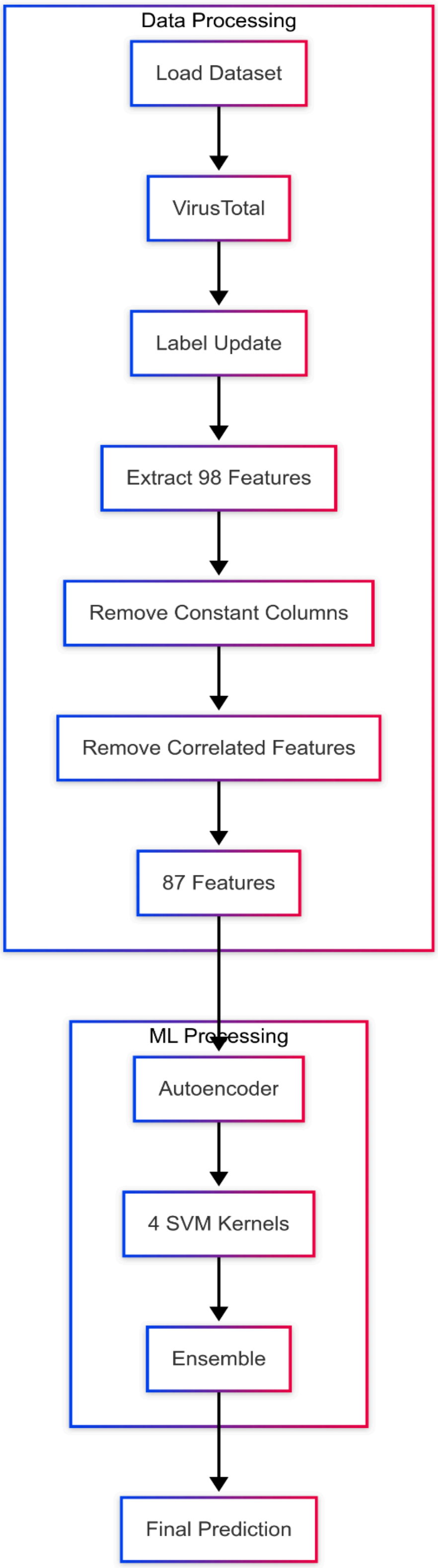
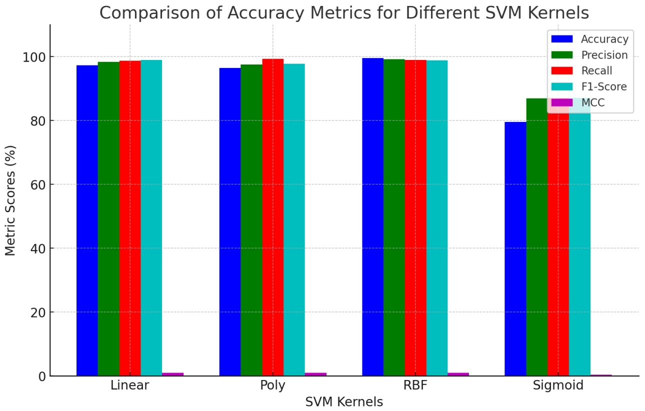
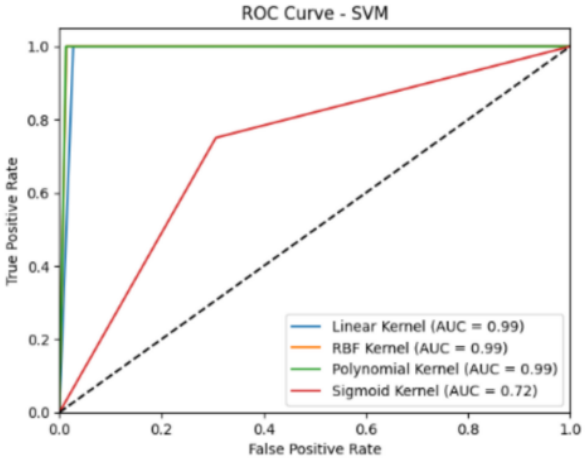


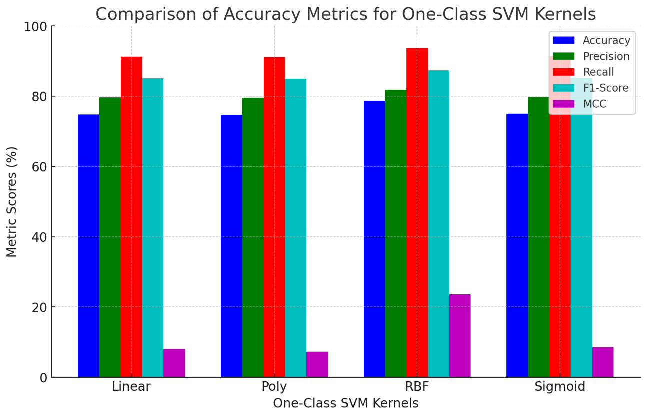
Fig: Block Diagram for Methodology

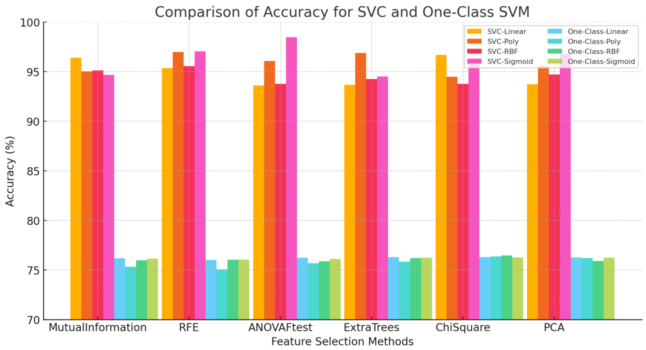
This study employs a multi-stage feature selection pipeline to optimize phishing URL detection using the url-phishing-extracted-features dataset. The dataset is first preprocessed by handling missing values (dropped if >30% missing) and normalizing features via MinMax scaling. Three primary feature selection techniques are applied: (1) Filter methods, including Pearson correlation (threshold=0.85) and mutual information (top-15 features selected), to eliminate redundant and low-variance features; (2) Wrapper methods, specifically recursive feature elimination (RFE) with a Random Forest classifier (10-fold CV, step size=5), to iteratively prune weakly contributing features; and (3) Embedded methods, such as L1-regularized logistic regression (C=0.01) and XGBoost feature importance (threshold=0.02), to leverage model-inherent selection. The reduced feature subset is evaluated using Logistic Regression, SVM-RBF, and Random Forest, with performance metrics (F1-score, AUC-ROC) validated through stratified 5-fold cross-validation.

Results







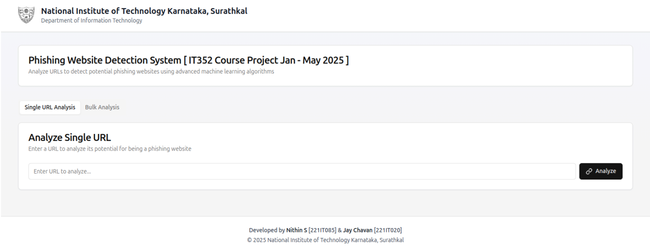


The RBF kernel generally performed the best, while the Sigmoid kernel was the weakest across most metrics. In one class SVM, although the scores are lower overall compared to standard SVM, the RBF kernel still showed relatively better performance. The metrics were quite similar across kernels with only small differences.

The SVC models achieved high accuracy (in the mid-90s percent range) consistently, regardless of the feature selection method used. The One-Class SVM models had lower accuracy (around the mid-70s percent range), again with only minor differences between feature selection methods.

RBF kernels tend to offer the best performance among the different kernels for both SVM and One-Class SVM. The SVC models clearly outperform the One-Class SVM models in terms of accuracy, precision, and other metrics, and feature selection method doesn't drastically change these results.

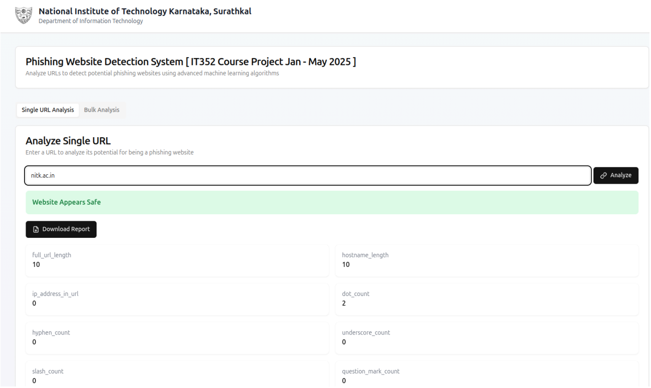
Web application


National Institute of Technology Karnataka, Surathkal
Department of Information Technology

Phishing Website Detection System [IT352 Course Project Jan - May 2025]
Analyze URLs to detect potential phishing websites using advanced machine learning algorithms

Analyze Single URL
Enter a URL to analyze its potential for being a phishing website

Developed by **Nithin S** [221IT085] & **Jay Chavan** [221IT020]
© 2025 National Institute of Technology Karnataka, Surathkal


National Institute of Technology Karnataka, Surathkal
Department of Information Technology

Phishing Website Detection System [IT352 Course Project Jan - May 2025]
Analyze URLs to detect potential phishing websites using advanced machine learning algorithms

Analyze Single URL
Enter a URL to analyze its potential for being a phishing website

nitk.ac.in

Website Appears Safe

full_url_length
10

hostname_length
10

ip_address_in_url
0

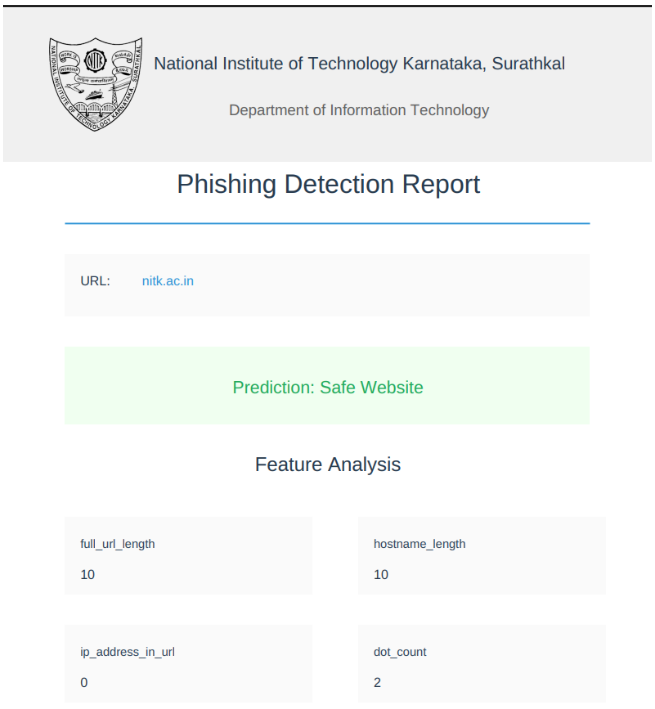
dot_count
2

hyphen_count
0

underscore_count
0

slash_count
0

question_mark_count
0


National Institute of Technology Karnataka, Surathkal
Department of Information Technology

Phishing Detection Report

URL: nitk.ac.in

Prediction: Safe Website

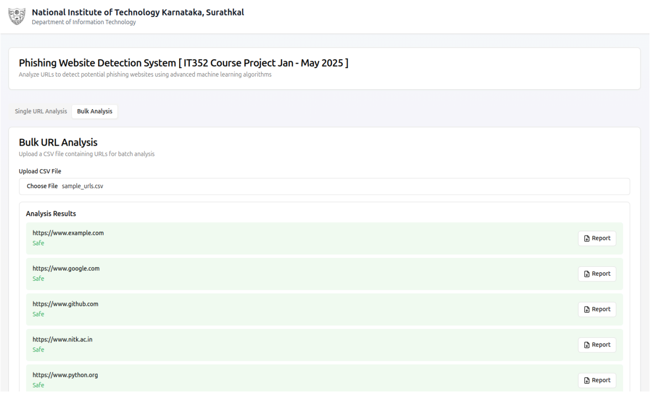
Feature Analysis

full_url_length
10

hostname_length
10

ip_address_in_url
0

dot_count
2


National Institute of Technology Karnataka, Surathkal
Department of Information Technology

Phishing Website Detection System [IT352 Course Project Jan - May 2025]
Analyze URLs to detect potential phishing websites using advanced machine learning algorithms

Bulk URL Analysis
Upload a CSV file containing URLs for batch analysis

Upload CSV File
Choose File sample_urls.csv

Analysis Results

https://www.example.com
Safe

https://www.google.com
Safe

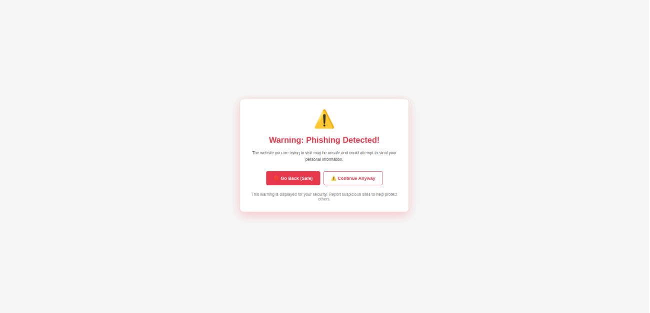
https://www.github.com
Safe

https://www.nitk.ac.in
Safe

https://www.python.org
Safe

We developed a Next.js + TypeScript + FastAPI application that analyzes URLs to predict whether they are phishing or legitimate. It utilizes a pickle file of our trained model to make predictions and displays all extracted features of the URL. The application also provides an option to download a detailed PDF report and supports bulk analysis by allowing users to upload a CSV file of URLs for evaluation.

Chrome Extension


⚠️

Warning: Phishing Detected!

The website you are trying to visit may be unsafe and could attempt to steal your personal information.

This warning is displayed for your security. Report suspicious sites to help protect others.

We have developed a Chrome Extension to seamlessly integrate our phishing detection model into everyday browsing. When a user clicks on a suspected phishing URL, the extension triggers a warning popup, providing options to either proceed or go back. If the user chooses to continue, the website is logged and exempted from future checks, ensuring a balance between security and convenience.

Conclusion

In conclusion, this project demonstrates a complete end-to-end solution for detecting phishing URLs. We started with a dataset of approximately 500,000 URLs, extracting 98 features from each. By applying autoencoders, we efficiently reduced these features to 15 key elements, simplifying the model without losing critical information. We then employed an SVM along with a majority voting classifier that combines predictions from four different kernels, ensuring robust and accurate detection. The integrated Next.js, TypeScript, and FastAPI application not only makes real-time predictions but also provides detailed feature analysis, downloadable PDF reports, and support for bulk URL analysis. This approach effectively balances sophisticated machine learning techniques with user-friendly, practical application design.ka

References

1. Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007). A comparison of machine learning techniques for phishing detection. In Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit (pp. 60-69).
2. Aljawarneh, S., Aldwairi, M., & Yassein, M. B. (2020). Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. Journal of Computational Science, 25, 152-160.
3. Almomani, A., Gupta, B. B., Atawneh, S., Meulenberg, A., & Almomani, E. (2018). A survey of phishing email filtering techniques. IEEE Communications Surveys & Tutorials, 15(4), 2070-2090.
4. Bahnsen, A. C., Bohorquez, E. C., Villegas, S., Vargas, J., & González, F. A. (2017). Classifying phishing URLs using recurrent neural networks. In Proceedings of the APWG Symposium on Electronic Crime Research (eCrime) (pp. 1-8).
5. Basnet, R. B., Sung, A. H., & Liu, Q. (2014). Feature selection for improved phishing detection. In Advanced Research in Applied Artificial Intelligence (pp. 252-261).
6. Feng, F., Zhou, Q., Shen, Z., Yang, X., Han, L., & Wang, J. Q. (2019). The application of a novel neural network in the detection of phishing websites. Journal of Ambient Intelligence and Humanized Computing, 10(1), 85-93.
7. Hannousse, A., & Yahiouche, S. (2021). Web page phishing detection: A survey. Journal of Network and Computer Applications, 167, 102731.
8. Jain, A. K., & Gupta, B. B. (2018). A machine learning based approach for phishing detection using hyperlinks information. Journal of Ambient Intelligence and Humanized Computing, 9(4), 1167-1178.
9. Kumar, V., Kumar, R., & Pandey, P. (2020). Phishing website detection using feature selection and classification techniques. International Journal of Computer Science and Information Security, 18(2), 51-58.
10. Lin, W., Xu, S., Yang, J., & Shi, L. (2018). Dimension reduction of web phishing detection data using PCA and fuzzy clustering. In IEEE International Conference on Intelligence and Security Informatics (pp. 197-202).
11. Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. Expert Systems with Applications, 117, 345-357.
12. Sahoo, D., Liu, C., & Hoi, S. C. (2017). Malicious URL detection using machine learning: A survey. arXiv preprint arXiv:1701.07179.
13. Tan, C. L., Chiew, K. L., & Sze, S. N. (2016). Phishing website detection using URL-assisted brand name weighting system. In International Symposium on Intelligence Computation and Applications (pp. 125-134).
14. Verma, R., & Dyer, K. (2015). On the character of phishing URLs: Accurate and robust statistical learning classifiers. In Proceedings of the 5th ACM Conference on Data and Application Security and Privacy (pp. 111-122).
15. Wei, W., Ke, Q., Nowak, J., Korytkowski, M., Scherer, R., & Woźniak, M. (2020). Accurate and fast URL phishing detector: A deep learning approach. Computer Networks, 178, 107275.
16. Zhu, E., Chen, Y., Ye, C., Li, X., & Liu, F. (2019). OFS-NN: An effective phishing websites detection model based on optimal feature selection and neural network. IEEE Access, 7, 73271-73284.