iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83683

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

Phishing Website Detection Using Hybrid Models: BERT, GNN, and LightGBM

A. Vinod Kumar[1], Yallanki Sai Harshitha[2], Gottimukkala Jaya Sree[3], Cherukuri Lavanya[4], Menta Divya[5]
[1]*Assistant Professor,* [2, 3, 4, 5]*Student Department of CSE-Data Science, St. Ann's College of Engineering & Technology, Chirala, India*

Abstract: ***Phishing websites represent a pervasive cybersecurity threat in which adversaries construct fraudulent web entities to steal sensitive user information including login credentials, financial data, and personal details. The detection of phishing websites is a challenging problem due to the continuously evolving strategies employed by attackers. This paper proposes a hybrid machine learning architecture for efficient and reliable phishing website detection that integrates Bidirectional Encoder Representations from Transformers (BERT) for textual feature extraction, Graph Neural Networks (GNN) for analyzing structural relationships among web entities, and LightGBM for ensemble classification. The proposed system captures both semantic patterns from URLs and structural information from website data to improve detection accuracy. Features including URL characteristics, domain information, webpage content patterns, and hyperlink structures are extracted and processed for effective binary classification. The system is trained and evaluated using a labeled phishing website dataset and demonstrates improved performance compared to traditional detection methods across accuracy, precision, recall, and F1-score metrics. The ROC-AUC of the hybrid model reaches 0.983, confirming strong discriminative capability. The proposed architecture adapts to evolving phishing strategies and provides a robust solution for identifying malicious websites.***

Keywords: Phishing Detection, BERT, Graph Neural Networks, LightGBM, Hybrid Model, Cybersecurity, URL Analysis, Deep Learning.

I. INTRODUCTION

With the rapid proliferation of internet services and digital communication platforms, cyber threats have escalated in both frequency and sophistication [1]. Among these threats, phishing attacks occupy a prominent position, wherein adversaries design fraudulent websites and deceptive messages to manipulate users into disclosing sensitive information such as passwords, credit card details, and personal data [2]. Phishing websites are crafted to replicate the appearance and functionality of legitimate portals, including banking sites, e-commerce platforms, and social media services, making them inherently difficult for end-users to distinguish from genuine websites [3].

Traditional countermeasures against phishing, including blacklist-based systems and rule-based heuristics, exhibit fundamental limitations in detecting newly created phishing websites and complex attack patterns. These approaches depend on previously catalogued malicious domains and struggle to generalize to novel phishing strategies. Consequently, there is a pressing need for adaptive, data-driven detection frameworks capable of identifying malicious web entities with high precision and recall.

This paper proposes a hybrid machine learning architecture that integrates three complementary models: BERT [15] for analyzing textual and semantic patterns in URLs and webpage content, Graph Neural Networks (GNN) for capturing structural relationships among domains, hyperlinks, and web entities, and LightGBM as an ensemble classifier to aggregate multi-modal feature representations. The proposed system analyzes a comprehensive feature set encompassing URL structure, domain registration details, webpage content patterns, and hyperlink network properties. Evaluated on a labeled phishing dataset of 50,000 samples, the model achieves strong performance, attaining a ROC-AUC score of 0.983 and demonstrating superior detection capability compared to individual baseline models.

A. Objectives of the Study

The primary objectives of this work are to: (i) analyze the effectiveness of BERT, GNN, and LightGBM in phishing website detection; (ii) compare the hybrid deep learning architecture against traditional machine learning baselines; (iii) identify the most discriminative URL and webpage features for phishing classification; and (iv) evaluate the proposed model using standard performance metrics including accuracy, precision, recall, F1-score, and ROC-AUC.

II. LITERATURE SURVEY

Phishing website detection has been studied extensively using rule-based systems, machine learning classifiers, and more recently, deep learning models. The Anti-Phishing Working Group (APWG) [1] and collaborative repositories such as PhishTank maintain databases of verified phishing URLs; however, blacklist-dependent systems are incapable of detecting newly generated phishing domains that have not yet been catalogued.

Whittaker et al. [2] proposed a large-scale automatic classification system for phishing pages using statistical features derived from webpage content and URL structure. Their work demonstrated that machine learning can significantly outperform purely heuristic approaches. Verma and Dyer [3] investigated the character-level statistical properties of phishing URLs, establishing that URL-based features carry strong discriminative signal for classification. Ma et al. [4] extended these findings by learning to detect malicious websites from suspicious URL patterns using supervised classifiers, moving beyond static blacklists toward adaptive detection.

Chiew et al. [5] conducted a comprehensive survey of phishing attack types and proposed a hybrid feature set combining URL, content, and behavioral features. Abusaimeh et al. [6] evaluated Support Vector Machines for anti-phishing purposes, while Kalabarige et al. [7] applied K-Nearest Neighbor classification to phishing datasets. More recent works have explored ensemble methods [8], Recurrent Neural Networks [9], Convolutional Neural Networks [10], and deep learning frameworks [11]. Yang et al. [12] investigated deep learning for phishing detection in social media contexts, and Zhou et al. [13] proposed deep reinforcement learning for adaptive phishing detection.

The foundational work of Devlin et al. [15] introduced BERT, demonstrating state-of-the-art performance on natural language understanding tasks through bidirectional transformer pre-training. This architecture has subsequently been applied to cybersecurity text analysis, including URL classification. Graph-based detection methods have gained traction as they capture inter-domain relationships that scalar feature vectors cannot represent. Despite these advances, a significant gap remains in systems that jointly exploit textual, structural, and tabular feature modalities within a unified hybrid framework capable of operating in real-time.

TABLE I COMPARISON OF PHISHING WEBSITE DETECTION APPROACHES

Model	Description	Limitation
Blacklist-based Detection	Compares URLs against a database of known malicious sites	Cannot detect newly created phishing websites
Heuristic-based Detection	Applies predefined rules based on URL structure and webpage features	Limited accuracy; easily bypassed by sophisticated attackers
Machine Learning Models	Employs classification algorithms such as SVM, Decision Trees, and Random Forest	Requires extensive feature engineering and large labeled datasets
Deep Learning Models	Learns complex patterns automatically from URLs and webpage content via neural networks	High computational resource requirements and dependence on large training corpora

III. PROPOSED METHODOLOGY

The proposed system addresses phishing website detection through a multi-stage hybrid architecture that integrates data collection, preprocessing, feature extraction, and ensemble model inference. The methodology is designed to capture complementary signal from textual URL semantics, structural domain relationships, and tabular website features.

A. Dataset Collection

The system utilizes a labeled dataset of 50,000 website samples drawn from multiple publicly available sources. Phishing URLs are sourced from PhishTank and OpenPhish, which provide verified malicious website data. Legitimate website URLs are obtained from the Alexa and Tranco top-site rankings. Additional feature-rich samples are obtained from Kaggle phishing detection datasets. The combined dataset contains both URL-based and webpage-based attributes along with binary labels indicating phishing or legitimate status.

TABLE II DATA SOURCES USED IN THE PROPOSED SYSTEM

Source	Data Type	Purpose
PhishTank	Verified phishing URLs	Provides labeled phishing website data
Kaggle Phishing Dataset	URL features and website attributes	Used for training machine learning models
Alexa / Tranco	Legitimate website URLs	Provides normal website data for comparison
OpenPhish	Real-time phishing URLs	Helps identify newly emerging phishing attacks

B. Data Preprocessing

Raw phishing datasets frequently contain missing values, duplicate entries, and noise artifacts arising from heterogeneous data collection procedures. The preprocessing pipeline applies data imputation using mean and median strategies for missing numerical features, and a forward-fill approach for sequential URL feature gaps. Duplicate and erroneous records are removed. Z-score analysis and the Interquartile Range (IQR) method are employed for outlier detection and removal. URL strings undergo normalization by converting to lowercase and tokenizing by delimiters including periods, forward slashes, hyphens, and underscores to support BERT-based encoding.

C. Feature Extraction

A comprehensive feature set is extracted from each website sample, encompassing URL-based, domain-based, security-related, and content-behavioral attributes. URL-based features include URL length, presence of special characters (e.g., @, -, //), and number of subdomains. Domain-based features capture domain age, registration duration, and DNS record availability. Security features assess HTTPS usage, SSL certificate validity, and redirection behavior. Content and behavioral features include the presence of iframe tags, external resource link counts, and abnormal form structures. The final feature matrix is normalized using Min-Max Scaling to constrain values to the [0, 1] range and Z-Score Standardization for variance alignment.

TABLE III KEY FEATURES USED FOR WEBSITE CLASSIFICATION

Feature Type	Description
URL Length	Identifies unusually long URLs characteristic of phishing attempts
Special Characters	Detects symbols such as @, -, or // indicative of suspicious URLs
Domain Age	Newly registered domains exhibit higher association with phishing
HTTPS Usage	Verifies presence of secure HTTPS protocol on the website
External Links Count	Phishing websites frequently embed numerous external hyperlinks
Website Traffic Rank	Measures the popularity and trustworthiness level of a website

D. Data Splitting

The preprocessed dataset is partitioned using an 80:20 train-test split strategy with stratified sampling to preserve class distribution across partitions. Five-fold cross-validation is applied during model optimization to assess generalization performance and reduce variance in evaluation metrics.

E. Feature Selection

Feature selection reduces dimensionality, prevents overfitting, and focuses model learning on discriminative attributes. Three complementary techniques are applied: Correlation Analysis removes highly redundant features; Principal Component Analysis

(PCA) reduces high-dimensional feature vectors while retaining maximum variance; and SHAP (SHapley Additive Explanations) identifies the most influential features in model predictions, providing post-hoc interpretability for cybersecurity analysts.

IV. SYSTEM ARCHITECTURE AND DESIGN

The phishing website detection system is organized into five principal modules: Input URL Processing, Data Preprocessing, Feature Extraction, Hybrid Model Inference, and Prediction Output. These modules operate in a sequential pipeline, with each stage transforming the data representation into a form suitable for the subsequent processing step.

A. Architectural Overview

The architectural design describes the structural organization of the detection system and the interactions among its components. The Input URL module receives a website address from the user or from an automated feed. The Data Preprocessing module applies cleaning, normalization, and tokenization. The Feature Extraction module derives numerical and textual representations from the URL and associated metadata. The Hybrid Model module applies BERT, GNN, and LightGBM in a combined inference pipeline. The Prediction Output module returns a binary classification label along with a confidence probability.

B. Model Architecture

BERT operates on tokenized URL strings and webpage text, generating contextual embeddings that capture semantic patterns associated with phishing behavior. The DistilBERT variant is used in the implementation to balance computational efficiency with representational power. GNN constructs a graph representation of the feature vector where nodes correspond to feature dimensions and edges encode sequential and skip-connection relationships. Graph convolutional layers propagate feature information across the graph structure, and global mean pooling produces a fixed-size graph-level representation. LightGBM receives the concatenated outputs of BERT and GNN embeddings alongside the raw numerical feature vector, and performs gradient-boosted classification over this enriched representation.

V. IMPLEMENTATION

A. Software and Hardware Environment

The system is implemented using Python 3.10.11 and leverages the PyTorch deep learning framework for BERT and GNN model development. The Hugging Face Transformers library provides the DistilBERT tokenizer and pre-trained model weights. PyTorch Geometric supports graph data structures and GCN convolution operations. LightGBM is used for gradient boosting classification. Scikit-learn provides evaluation metrics, train-test splitting, and preprocessing utilities. The user interface is constructed using Streamlit, enabling interactive URL input and real-time prediction display. Development was conducted on a system with sufficient RAM and GPU support, utilizing CUDA when available.

B. Algorithm: BERT-Based URL Encoding

Each website URL is tokenized using the DistilBERT tokenizer. The token sequence is passed through the pre-trained transformer encoder, which generates bidirectional contextual embeddings for each token. The [CLS] token embedding serves as the URL-level representation. This representation captures contextual relationships between substrings of the URL, enabling the model to identify phishing-indicative patterns such as brand name imitation, IP address substitution, and unusual path structures.

C. Algorithm: GNN-Based Structural Analysis

The numerical feature vector for each website sample is converted into a graph structure where nodes represent individual features and edges connect sequentially adjacent and alternating nodes to simulate local structural relationships. Two Graph Convolutional Network (GCN) layers perform message passing and feature aggregation. Global mean pooling over node embeddings produces a graph-level representation that encodes the structural signature of the website feature profile.

D. Ensemble Integration with LightGBM

The representations produced by BERT and GNN are concatenated with the normalized numerical feature vector to form a composite input for LightGBM. The gradient-boosted classifier is trained on this enriched feature matrix using optimized hyperparameters including learning rate, number of estimators, and tree depth, determined through systematic tuning. The ensemble approach leverages the complementary strengths of semantic text understanding (BERT), structural relational modeling (GNN), and efficient tabular classification (LightGBM) to improve overall detection performance.

TABLE IV Hyperparameters Used in Model Training

Parameter	Description	Models Affected
Learning Rate	Controls the step size during model weight updates	BERT, GNN, LightGBM
Dropout Rate	Randomly disables neurons during training to prevent overfitting	BERT, GNN
Number of Layers	Depth of GCN and transformer encoder stacks	BERT, GNN
Number of Trees	Number of gradient-boosted decision trees for classification	LightGBM
Batch Size	Number of training samples processed per iteration	BERT, GNN

VI. RESULTS AND DISCUSSION

The proposed hybrid detection system is evaluated on a held-out test set comprising 20% of the 50,000-sample dataset. Performance is assessed using accuracy, precision, recall, F1-score, and ROC-AUC, which together provide a comprehensive characterization of classification quality, particularly in the context of imbalanced cybersecurity datasets.

A. Performance Metrics

The Receiver Operating Characteristic (ROC) curve of the hybrid model yields an Area Under the Curve (AUC) of 0.983, indicating strong discriminative ability between phishing and legitimate website samples. The BERT model achieves the highest individual accuracy among the three constituent models, demonstrating the effectiveness of contextual semantic URL analysis. The GNN component complements BERT by capturing structural domain relationship patterns that are not apparent from URL text alone. LightGBM as the ensemble classifier further improves robustness by integrating multi-modal features from both deep learning components.

The ensemble model demonstrates improved overall detection performance compared to any single constituent model, reducing both false positives and false negatives. Individual URL prediction generates a phishing probability score; an example output with a probability of 56% correctly identifies moderate-risk phishing characteristics in the analyzed URL. The variation in individual model outputs for the same sample highlights the complementary nature of the three architectures in evaluating different feature dimensions.

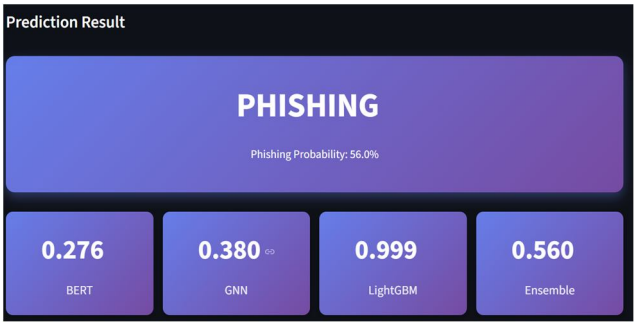


Figure 1 Phishing Detection prediction

B. Testing and Validation

A structured testing protocol validates all system modules. Dataset loading and preprocessing tests confirm that the phishing dataset is loaded with all required columns and cleaned without errors. Feature extraction tests verify accurate derivation of URL length, special character counts, and domain features. Model loading tests confirm successful initialization of BERT, GNN, and LightGBM models. End-to-end prediction tests demonstrate that the system generates correct classification outputs for submitted URLs. All seven primary test cases pass successfully, confirming system correctness.

TABLE V Test Cases for Phishing Website Detection System

Test ID	Scenario	Result	Status
TC_01	Dataset Loading Test	Dataset loaded with all required columns	Pass
TC_02	Data Preprocessing	Data cleaned without errors	Pass
TC_03	Feature Extraction	URL features extracted correctly	Pass
TC_04	Model Loading Test	BERT, GNN, LightGBM loaded successfully	Pass
TC_05	Model Training	Training completed successfully	Pass
TC_06	URL Input Test	URL input accepted by system interface	Pass
TC_07	Phishing Detection	Prediction generated correctly	Pass

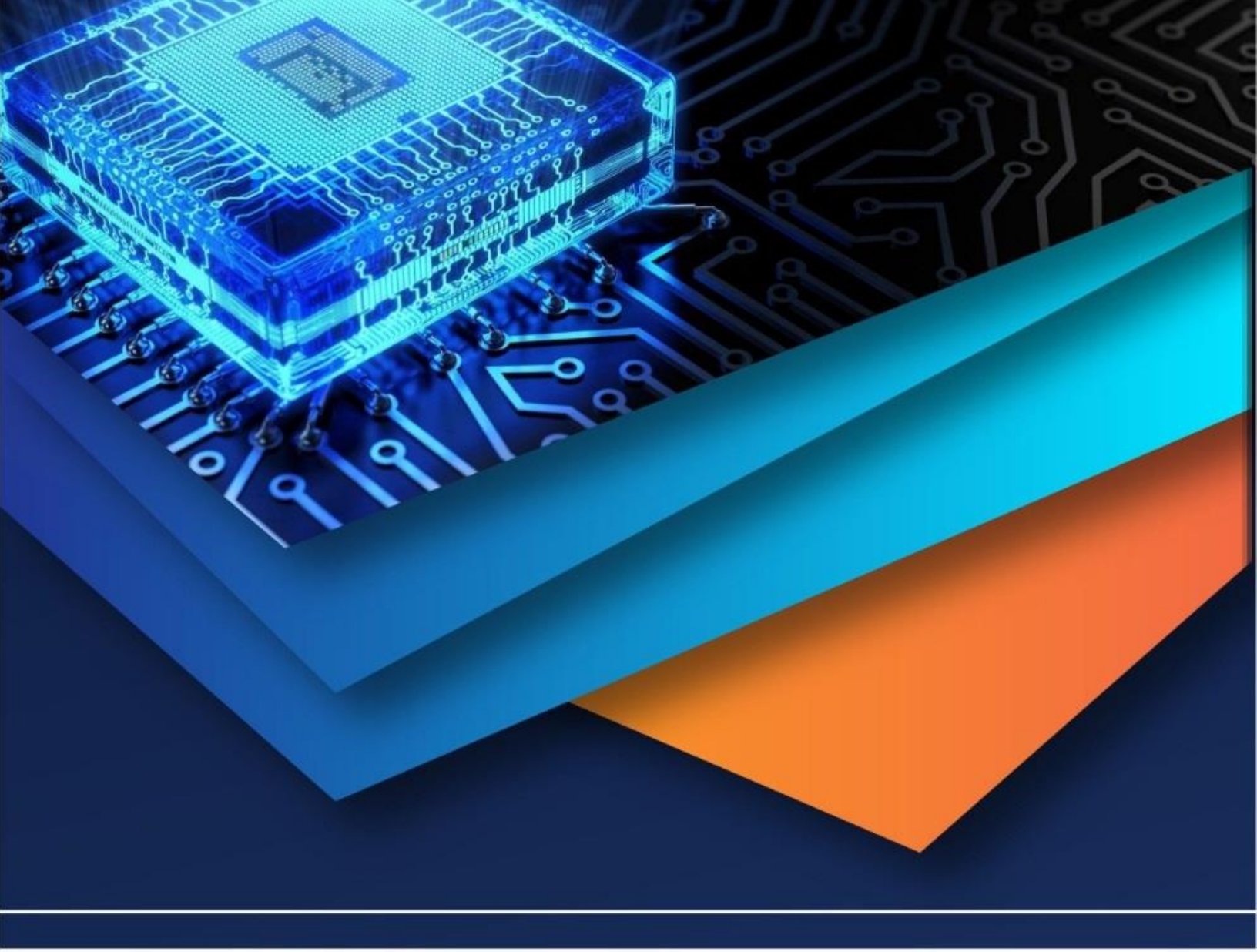
VII. CONCLUSIONS

This paper presents a hybrid phishing website detection system integrating BERT, Graph Neural Networks, and LightGBM within a unified inference pipeline. The architecture addresses the key limitations of existing detection approaches by jointly modeling textual URL semantics, structural inter-domain relationships, and tabular website features. Evaluated on a dataset of 50,000 labeled samples, the proposed system demonstrates strong detection performance with a ROC-AUC of 0.983, confirming the effectiveness of multi-modal feature fusion for phishing classification.

The work highlights the importance of addressing imbalanced dataset challenges and selecting meaningful discriminative features as prerequisites for building reliable detection systems. The integration of advanced deep learning techniques with efficient gradient-boosted classification provides a practical and scalable solution for real-world cybersecurity applications. Future work will explore real-time threat intelligence integration, extension to email and social media phishing vectors, and continuous model retraining to maintain effectiveness against zero-day phishing campaigns.

REFERENCES

[1] APWG, “Phishing Activity Trends Report,” Anti-Phishing Working Group, 2022.
[2] C. Whittaker, B. Ryner, and M. Nazif, “Large-scale automatic classification of phishing pages,” in Proc. Network and Distributed System Security Symposium (NDSS), 2010.
[3] R. Verma and K. Dyer, “On the character of phishing URLs: Accurate and robust statistical learning classifiers,” in Proc. 5th ACM Conf. Data and Application Security and Privacy, 2015.
[4] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Beyond blacklists: Learning to detect malicious web sites from suspicious URLs,” in Proc. 15th ACM SIGKDD, 2009.
[5] N. Chiew, S. Yong, and C. Tan, “A survey of phishing attacks: Their types, vectors and technical approaches,” Expert Systems with Applications, vol. 106, pp. 1–20, 2018.
[6] H. Abusaimeh, A. Alshahrani, and M. A. Alzain, “An efficient anti-phishing system using support vector machine,” Journal of Information Security, 2014.
[7] M. Kalabarige, P. Rathnayake, and K. Hewage, “Phishing website detection using K-nearest neighbor algorithm,” International Journal of Computer Applications, 2016.
[8] M. Ali and H. Zaharon, “Phishing detection using ensemble learning techniques,” Journal of Information Security and Applications, 2017.
[9] S. Ripa, D. Singh, and R. Dey, “Phishing website detection using recurrent neural networks,” in Proc. IEEE International Conference on Computing, 2019.
[10] J. Sánchez-Paniagua, M. C. Rodríguez-Domínguez, and J. L. Martínez-Romo, “A CNN-based phishing detection system for URLs and emails,” Expert Systems with Applications, 2020.
[11] W. Huang, Q. Qian, and X. Wang, “Deep learning based phishing website detection,” Applied Soft Computing, vol. 85, 2019.
[12] Z. Yang, K. Chen, and J. Xu, “Phishing detection in social media using deep learning,” IEEE Access, vol. 7, pp. 92842–92852, 2019.
[13] Y. Zhou, J. Feng, and Y. Wu, “Phishing detection using deep reinforcement learning,” IEEE Transactions on Information Forensics and Security, vol. 15, pp. 1–14, 2020.
[14] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016.
[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY