IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14803

PhishHybridNet: A Multi-Modal Deep Learning and Ensemble Approach for Robust Phishing URL Detection

Nagesha N M[1], Dr.Prabha R[2], Prof. Veena Potdar[3]

Student, M.tech Department of Computer Science & Engineering, Dr. Ambedkar Institute of Technology, Bengaluru[1]

Professor, Department of Computer Science & Engineering, Dr. Ambedkar Institute of Technology, Bengaluru[2]

Associate Professor, Department of Computer Science & Engineering, Dr. Ambedkar Institute of Technology, Bengaluru[3]

Abstract: Phishing attacks pose a serious cyber security threat by imitating legitimate websites to steal sensitive data. This study presents a hybrid phishing detection system integrating Machine Learning (ML), Deep Learning (DL), and Ensemble Learning (EL). Feature selection techniques such as Information Gain, Gain Ratio, and Principle component Analysis (PCA) are applied to extract the most relevant indicators from a dataset of 11,055 URLs. ML classifiers (SVM, DT, KNN), EL models (RF, XGBoost, AdaBoost), and DL architectures (LSTM, GRU, CNN) are used. A hybrid model fuses LSTM and GRU outputs, processed by ensemble classifiers and finalized by a meta-classifier. The model captures both structural and sequential URL features, improving accuracy, reducing false positives, and enabling real-time adaptability. The framework can be deployed in email clients, browsers, or gateways to safeguard users from phishing threats. This scalable and intelligent system outperforms individual models and adapts to evolving phishing tactics, contributing to a more secure online ecosystem.

Keywords: Phishing, Machine Learning, Deep Learning, Ensemble Learning, Hybrid Model, Cyber security

I. INTRODUCTION

This research highlights the urgent need for advanced techniques to effectively detect phishing websites. To address this, we explore the use of cutting-edge algorithms from machine learning, ensemble learning, and deep learning domains. Phishing attacks pose a serious cyber security risk, often leading to data theft and financial loss. The proposed system employs feature selection techniques like Information Gain, Gain Ratio, and Principal Component Analysis to enhance the prediction of whether a website is phishing or legitimate. The model is trained on a dataset containing 11,055 website samples. Results demonstrate that combining deep learning with ensemble learning leads to higher accuracy and better adaptability to new phishing methods.

Phishing involves tricking users with fake websites that closely imitate trusted ones. Attackers set up counterfeit sites and wait for victims to unknowingly submit sensitive information. Figure 1 illustrates the entire phishing lifecycle, showing how attackers deceive users. These phishing sites often replicate platforms like Facebook, Google, or Twitter, even imitating HTTPS and green lock symbols to appear trustworthy. This makes it difficult for users to distinguish between real and fake websites, pushing researchers to develop more intelligent detection systems.

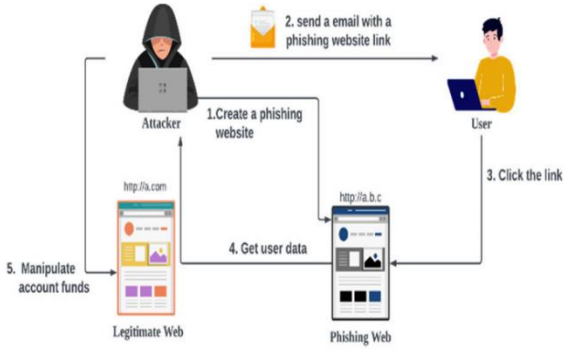

Figure 1: Life cycle of phishing

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14803

Figure 2 explains URL structure using HTTP protocol. Label 1 represents HTTP, which is used to request resources. Label 2 is the hostname, split into top-level domains and subdomains. Label 6 shows the server directory path, and label 7, such as 'v=AbcdEffGhIJ,' represents query parameters. URLs act as digital addresses, and HTTP facilitates communication between web clients and servers.

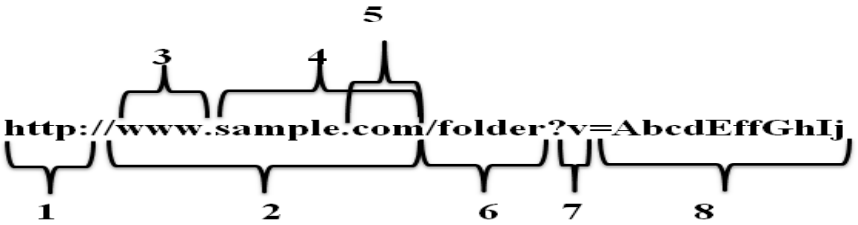

Figure 2: URL presentation based on HTTP

II. LITERATURE REVIEW

[1] Phishing has grown into a major cyber security threat, targeting both individuals and organizations. As phishing methods evolve, there is a growing need for advanced detection systems. Traditional heuristic-based detection methods have become less effective in combating these sophisticated attacks. With recent advancements, machine learning and deep learning offer more promising solutions. Algorithms such as Support Vector Machines (SVM), Decision Trees (DT), and K-Nearest Neighbors (KNN) are widely used for classifying phishing websites based on features like URL structure, page content, and user behavior. Additionally, ensemble techniques like Random Forest (RF) and AdaBoost enhance detection accuracy by combining multiple classifiers. Techniques like Information Gain (IG) and Gain Ratio (GR) are employed for selecting the most informative features, further improving performance.
[2] This study highlights the use of deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), for phishing detection. CNNs are highly effective in identifying phishing through image-based analysis, while RNNs and Long Short-Term Memory (LSTM) networks are suitable for analyzing sequential data such as browsing patterns or URL sequences. These models can learn intricate patterns from large datasets, enabling highly accurate detection of phishing threats. Multiple experiments demonstrate the success of these models in real-world cyber security applications.

[3] The researchers focused on comparing deep learning and traditional ML approaches. They reported that models like LSTM significantly outperform traditional methods in phishing detection tasks. One of the key advantages of deep learning highlighted in this work is its ability to automatically learn features from raw input data, thereby reducing dependency on manual feature engineering and simplifying the development process.

[4] Despite recent progress in phishing detection, several challenges remain most notably the dynamic and ever-changing nature of phishing strategies. The study emphasizes the difficulty of interpreting complex deep learning models, which can be a hurdle for cyber security professionals who need transparency in decision-making. To address these challenges, the researchers propose hybrid models that combine traditional ML methods with deep learning architectures, resulting in improved accuracy and adaptability.

[5] This study supports the idea of enhancing phishing detection by integrating feature selection techniques with hybrid architectures. The authors argue that such a combination leads to more robust models capable of addressing evolving phishing tactics. The paper emphasizes that future research should focus on real-time detection capabilities, efficient model optimization, and feature selection strategies to improve system performance. Hybrid models that fuse ML and DL are seen as a promising path forward in the fight against cyber threats.

III. METHODOLOGY

This project utilizes a multi-stage methodology that combines machine learning and deep learning techniques to build a hybrid phishing detection model.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14803

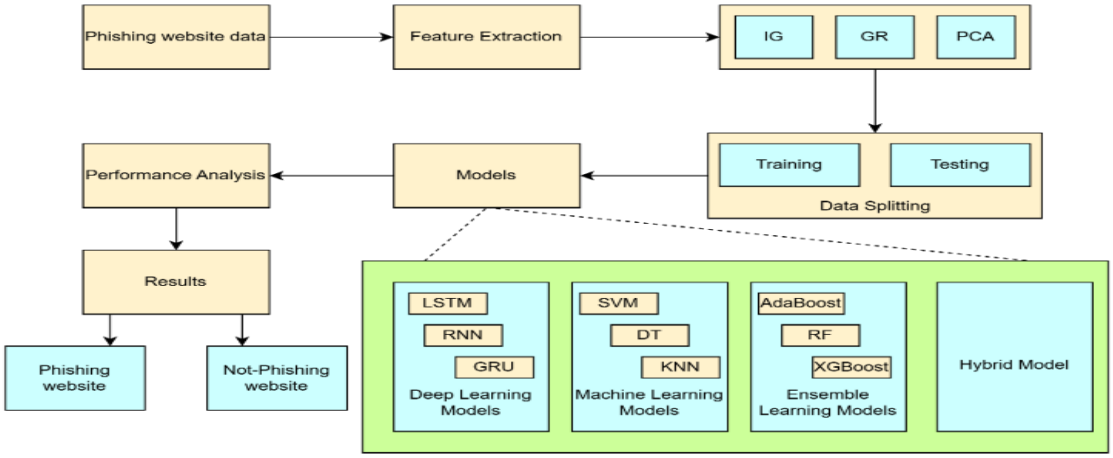

Figure 3: Architecture of the Proposed System

The system design begins with collecting and pre-processing data, followed by feature engineering and dimensionality reduction. These steps ensure that the input data is clean, relevant, and optimized for the models to work efficiently. Initially, a dataset containing website information such as URLs, domain attributes, and SSL certificate status is collected. The data undergoes extensive pre-processing to remove inconsistencies, missing values, and redundant entries, and to transform features into suitable formats for model training.

Feature engineering plays a critical role in enhancing the performance of the system. Here, RFM (Recency, Frequency, and Monetary) analysis is not applicable since the project deals with phishing detection; instead, domain-based and structural features of URLs are extracted. Dimensionality reduction is performed using Principal Component Analysis (PCA) to retain essential features while reducing the computational burden, which helps in visualizing and clustering website behaviours.

Next, customer segmentation through K-Means clustering, though originally intended for recommendation systems, is replaced here by feature clustering to group similar phishing behaviours. Then, a Convolutional Neural Network (CNN) model, modified for sequence analysis, is used to process URL feature matrices and detect phishing patterns. This model learns to extract relevant features and classify URLs using deep learning layers. The evaluation of the system includes various metrics such as confusion matrix, precision, recall, F1-score, and ROC curves, all of which contribute to assessing the reliability and accuracy of the proposed solution.

IV. ANALYSIS AND RESULTS

The result analysis of the phishing website detection system provides critical insights into the dataset characteristics, model performance, and evaluation outcomes. The analysis begins by examining the dataset itself, followed by interpreting the results of various machine learning, deep learning, and hybrid models implemented throughout the project.

Figure 4 illustrates the class distribution of the dataset, showing the count of samples belonging to phishing and legitimate categories. This visualization is crucial for assessing class imbalance, which can significantly impact the model's performance. A balanced dataset ensures that the model learns equally from both classes, whereas an imbalanced dataset may lead to biased predictions favoring the dominant class. Understanding this distribution helps in deciding whether sampling techniques or weighted models are necessary.

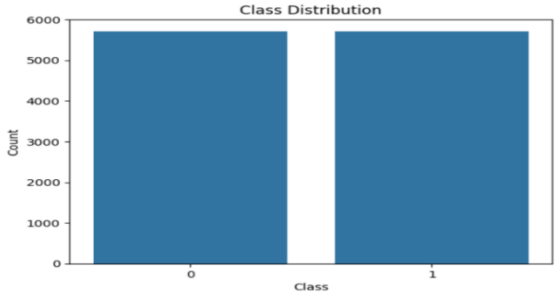

Figure 4: Class Distribution plot

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14803

Figures 5 and 6 demonstrate the confusion matrices for the Support Vector Machine (SVM) and Decision Tree (DT) models, respectively. These matrices help to evaluate how accurately each algorithm classifies the URLs. The SVM model tends to perform well in high-dimensional spaces, while DT offers interpretability but may suffer from overfitting.

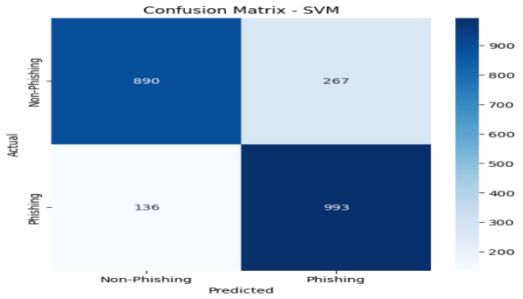

Figure 5: Confusion Matrix of SVM

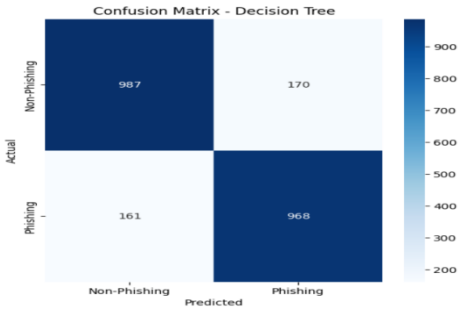

Figure 6: Confusion Matrix of DT

Figures 7 and 8 depict the results from K-Nearest Neighbors (KNN) and Random Forest (RF) classifiers. KNN, being a distance-based classifier, depends heavily on feature scaling and performs well with clean datasets. RF, on the other hand, is an ensemble method that reduces variance and often provides more stable results.

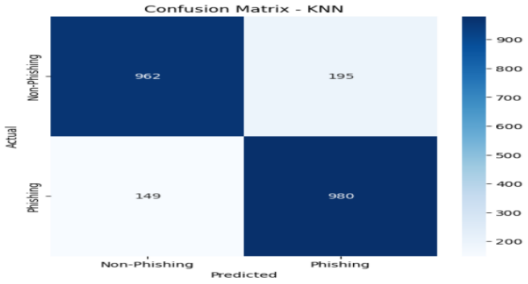

Figure 7: Confusion Matrix of KNN

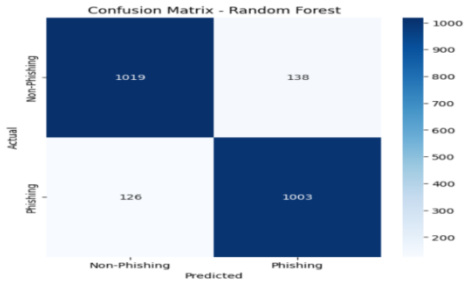

Figure 8: Confusion Matrix of RF

Figures 9 and 10 display the performance of AdaBoost and XGBoost classifiers. AdaBoost iteratively adjusts the weights of data points to minimize classification error, while XGBoost is a more regularized boosting algorithm designed for improved speed and performance.

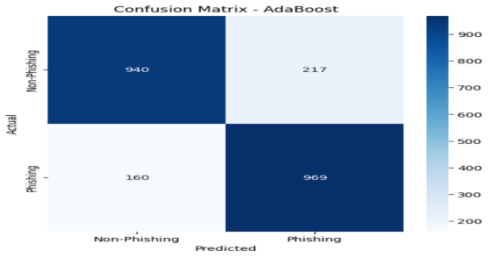

Figure 9: Confusion Matrix of AdaBoost

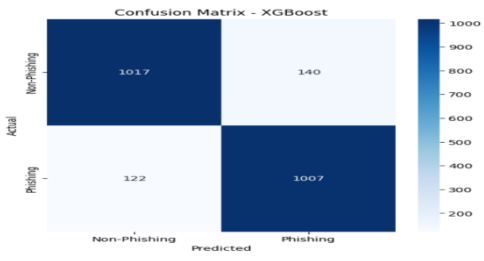

Figure 10: Confusion Matrix of XGBoost

Figure 11 through Figure 13 show the performance of deep learning models—LSTM, GRU, and RNN. These models are particularly effective in capturing sequential patterns and dependencies within the data. The LSTM model excels at learning long-term relationships, whereas GRU is a lighter and faster variant that maintains comparable performance. The standard RNN suffers from vanishing gradient problems but still serves as a baseline deep learning approach. The confusion matrices reveal how effectively these models distinguish between phishing and legitimate websites, based on temporal and sequential features.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14803

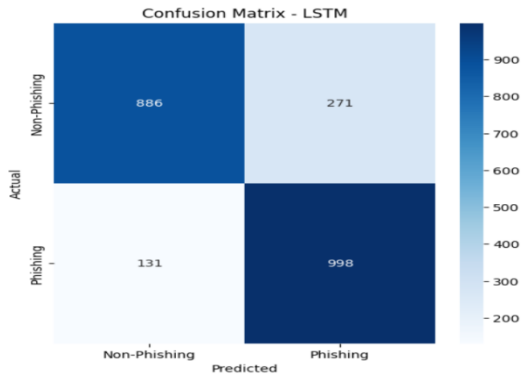

Figure 11: Confusion Matrix of LSTM

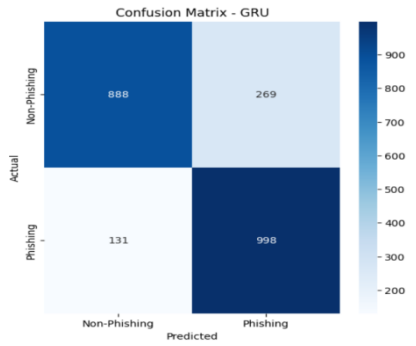

Figure 12: Confusion Matrix of GRU

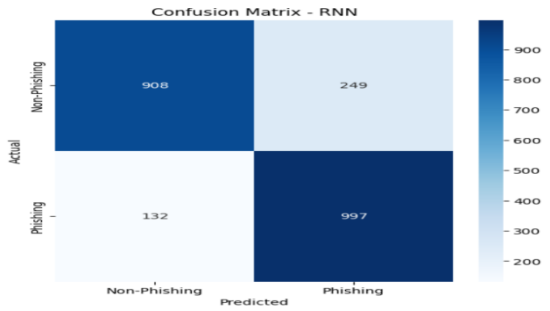

Figure 13: Confusion Matrix of RNN

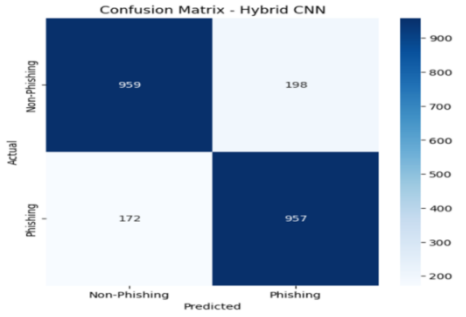

Figure 14: Confusion matrix of Hybrid model

In Figure 14, the hybrid model is evaluated using a CNN integrated with deep learning outputs. This model combines the strengths of convolutional feature extraction and sequential modeling, yielding high accuracy and better generalization. The confusion matrix of the hybrid model shows a low rate of false positives and false negatives, indicating its reliability in real-time phishing detection scenarios.

V. CONCLUSION AND FUTURE WORK

In conclusion, this project introduces a hybrid phishing website detection system that combines traditional machine learning techniques with advanced deep learning models to enhance detection accuracy and reliability. By applying effective feature selection methods such as Information Gain and Gain Ratio, the model focuses on the most impactful attributes, improving classification efficiency while reducing complexity. The integration of models like LSTM, GRU, Random Forest, and XGBoost allows the system to leverage both sequential behavioral patterns and ensemble decision-making, resulting in high accuracy and robust performance across multiple evaluation metrics.

To keep up with the evolving tactics of cybercriminals, the proposed system can be extended in several meaningful directions. First, real-time data collection and live URL scanning can be implemented to ensure that the system remains responsive to new types of phishing websites as they emerge. Integration with web browser plugins or email filtering systems can bring this detection mechanism closer to end-users. Second, incorporating Natural Language Processing (NLP) techniques could enhance detection by analyzing the textual content of websites or email messages for suspicious patterns. Additionally, expanding the model to handle multilingual phishing attempts and image-based attacks would broaden its applicability. system.

REFERENCES

[1]. Umezara, Kashif Ayyub, Hikmat Ullah Khan, Ali Daud, Tariq Alsahfi , And Saima Gulzar Ahmad, “Phishing Website Detection Using Deep Learning Models”, Proceedings of the IEEE Access, 0.1109/ACCESS.2024.3486462, 26 August 2024.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14803

[2]. Abdul Karim, Samir Brahim Belhaouari, Mobeen Shahroz, Khabib Mustofa, And S. Ramanakumarjoga, "Phishing Detection System Through Hybrid Machine Learning Based on URL" Proceedings of the IEEE Access, 10.1109/ACCESS.2023.3252366, 18 April 2023.

[3]. Ozgur Koray Sahingoz, Ebubekir Buber, Andeminkugu, "DEPHIDES: Deep Learning Based Phishing Detection System", Proceesings of the IEEE Access, 10.1109/ACCESS.2024.3352629, December 2023.

[4]. Adarsh Mandadi, Saikiran Boppan, Vishnu Ravell, Prof. Dr R Kavitha, "Phishing Website Detection Using Machine Learning", Proceedings of the IEEE 7th International conference for Convergence in Technology (I2CT) Pune, India. Apr 07-09, 2022.

[5]. Fatima Salahdine, Zakaria El Mrabet, Naima Kaabouch, "Phishing Attacks Detection A Machine Learning-Based Approach", Proceedings of the 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2021.

[6]. Michael A. Ivanov, Bogdana V. Kliuchnikova, Ilya V.Chugunkov, Anna M. Plaksina, "Phishing Attacks and Protection Against Them", Proceedings of the 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus),2021.

[7]. Ishita Saha, Dhiman Sarma, Rana Joyti Chakma, Mohammad Nazmul Alam, "Phishing Attacks Detection using Deep Learning Approach", Proceedings of the Third International Conference on Smart Systems and Inventive Technology (ICSSIT 2020), 2020.

[8]. Ayman El Aassal, Shahryar Baki, Avisha Das, And Rakesh M. Verma, "An In-Depth Benchmarking and Evaluation of Phishing Detection Research for Security Needs", Proceedings of the IEEE Access, 10.1109/ACCESS.2020.2969780, 2020.

[9]. Abdul Razaque, Mohamed Ben Haj Frej, Dauren Sabyrov, Aidana Shaikhyn, "Detection of Phishing Websites using Machine Learning", Proceedings of the IEEECloudSummit48914. 2020. 00022, 2020.