IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, March 2026
Impact Factor: 8.2

An Intelligent Hybrid Framework for Phishing Website Detection Using Feature Selection and Deep Learning Models

Rahul and Pratap Singh Patwal
Department of Computer Science and Engineering
Laxmi Devi Institute of Engineering & Technology, Alwar
mrahulcse9@gmail.com and pratappatwal@gmail.com

Abstract: *Phishing is still a significant cybersecurity issue, in which criminals create fake websites to deceive users into providing private information like financial details or passwords. In this study, an intelligent hybrid phishing detection system based on feature selection techniques and advanced models of machine learning and deep learning is proposed. Advanced methods such as Chi-Square, Information Gain and Recursive Feature Elimination improve feature optimization and enhance model performance. Random Forest, XGBoost, GCN, and TabTransformer models exhibit higher accuracy, precision, recall, and F1-scores, providing robust and scalable phishing website detection*

Keywords: Phishing Detection, Cybersecurity, Machine Learning, Deep Learning, Feature Selection, Graph Convolutional Network, TabTransformer

I. INTRODUCTION

Internet technologies and web communication have seen huge strides forward and online communication has become a major part of everyone's daily lives, which has added to the cybersecurity threats facing the world. Phishing attacks, one of the most dangerous cybercrimes
against individuals, organizations and financial institutions, is one of these threats. Phishing involves the creation of fake websites or emails that look like a real one to deceive the user into providing information, including passwords, banking information and personal details.

The traditional ways to detect phishing consist primarily in blacklist systems, signature-based detection and manual verification. But these techniques will not work on new phishing sites because the attacker is constantly changing the URL, web page structure and domain characteristics. With the increasing sophistication of phishing attacks, intelligent automated detection systems are needed to enhance cybersecurity protection.

The application of machine learning techniques has been found to be effective in the detection of phishing websites by the detection of hidden patterns and suspicious characteristics of the website. Recently deep learning model became popular in the area of implementing complex relationship based in big data without manual feature engineering. For structured and relational data, advanced models like Graph Convolutional Networks (GCN) and TabTransformer can effectively handle such data to achieve better classification results.

In this study, a hybrid intelligent phishing detection framework is proposed by combining feature selection methods, machine learning and deep learning models. The objective of the framework is to increase the accuracy of the classification of phishing websites, decrease false positive and increase the detection efficiency.

II. LITERATURE REVIEW

There are several research works that have investigated the use of machine learning or deep learning techniques to detect phishing websites. The traditional machine learning algorithms, including Decision Tree, Support Vector

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 8, March 2026 Impact Factor: 8.2

Machine (SVM), Naïve Bayes, and Random Forest are widely used for phishing website identification based on various features such as URL structure, domain age, SSL certificates, and webpage content. Conventional models were outperformed by the ensemble models like Random Forest and XGBoost in terms of classification accuracy and robustness. In recent years, the introduction of deep learning models, including CNN, RNN, and LSTM, has made it possible to automatically identify intricate phishing patterns from the characteristics of websites. In addition, Graph Convolutional Networks (GCN) are used to understand the relationships between websites and hyperlinks, and TabTransformer is used to enhance the classification of structured data with contextual embeddings. Many of existing studies, however, do not have an optimized feature selection technique to reduce the complexity of the computation. Thus, the combination of feature selection with state-of-the-art deep learning models continues to be a crucial research challenge for phishing detection systems.

III. PROBLEM STATEMENT

There are some limitations in the existing phishing detection systems that affect their ability to work in real-world environments. Traditional blacklist-based systems cannot identify sites that have just been created as they must be previously known phishing sites. High dimensional data sets which contain redundant and irrelevant features also cause the machines to be less effective when implementing many machine learning algorithms.

One of the other challenges is the Evolution of Phishing Attacks. The attackers constantly change the domain architecture, page layouts and phishing techniques to escape the detection systems. The other existing models also have high false positive rates, misclassifying legitimate websites as phishing websites.

Deep learning models can enhance phishing detection accuracy but may consume a lot of computing resources and be lacking of feature optimization mechanisms. Thus, there is a need for an intelligent hybrid framework that integrates the feature selection technique with machine learning and deep learning models to enhance the accuracy, scalability and efficiency of phishing detection.

IV. PROPOSED METHODOLOGY

4.1 Framework Overview

A smart hybrid phishing detection system based on feature selection methods and machine learning/deep learning models is proposed. The entire framework can be broken down into several stages: data collection, data preprocessing, feature selection, model training, and performance evaluation.

4.2 Dataset Collection

A publicly available data set of phishing websites from cybersecurity repositories and Kaggle datasets are used in the study. This data set has a range of attributes, including details of the URL, domain, HTML structure and security features, for both valid and phishing website instances. The classification label is taken from the “status” field.

4.3 Data Preprocessing

Data pre-processing is carried out to enhance the quality of the data set and eliminate inconsistencies.

The pre-processing steps are:

Handling missing values
Removing duplicate records
Data normalization
Encoding categorical variables
Balancing class distribution

The preprocessing methods enhance the efficiency of the model and decrease noise in the data sets.

4.4 Feature Selection

Feature selection techniques are used to select the most relevant phishing features and remove features that are redundant. These methods are:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, March 2026
Impact Factor: 8.2

Chi-Square Test

The Chi-Square method is used to determine the statistical relationship between features and target labels.

Information Gain

Information Gain is a measure of the importance of features using entropy reduction.

It is also known as Recursive Feature Elimination (RFE).

RFE works by progressively eliminating less important features and creates a model that performs better.

The features chosen enhance the computational efficiency and minimize the complexity of the model.

4.5 Machine Learning Models

These are the following machine learning models implemented:

Decision Tree

Random Forest

Support Vector Machine (SVM)

XGBoost

Only the best features are used for training and evaluation of these models.

4.6 Deep Learning Models

Graph Convolutional Network (GCN): GCN applies graph based learning methods to analyze relationships between websites, domain structures and URLs.

TabTransformer: TabTransformer is a transformer-based contextual embedding model for structured tabular data that supports better phishing classification.

V. PROPOSED ARCHITECTURE

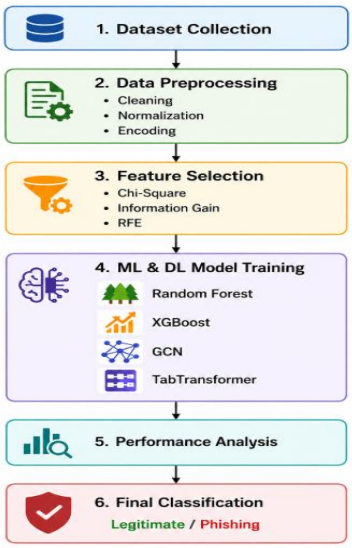

Figure: Hybrid Intelligent Framework for Phishing Website Detection Using Machine Learning and Deep Learning

VI. EXPERIMENTAL RESULTS

Various machine learning and deep learning models were tested on the proposed framework. Common metrics such as accuracy, precision, recall and F1-score were employed for performance analysis.

6.1 Performance Metrics

Accuracy: Measures percentage of classified websites that are correct.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 8, March 2026 Impact Factor: 8.2

Precision: Calculates the percentage of successfully predicted phishing sites.
Recall: Evaluates the correctness of a model in recognizing phishing websites.
F1-Score: Reflects the harmonic average of precision and recall.

6.2 Performance Comparison

Table: Performance Comparison of Machine Learning and Deep Learning Models for Phishing Website Detection

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	91.2%	90.5%	89.8%	90.1%
Random Forest	95.4%	94.8%	95.1%	94.9%
XGBoost	96.1%	95.7%	95.9%	95.8%
GCN	97.3%	97.0%	96.8%	96.9%
TabTransformer	98.2%	98.0%	97.9%	97.9%

Based on the results, it is observed that TabTransformer outperforms all the implemented models.

VII. DISCUSSION

The experimental results show that incorporating feature selection methods with deep learning models considerably enhance the phishing detection performance. The traditional machine learning models could reach acceptable classification accuracy, while the advanced deep learning models (GCN and TabTransformer) outperformed traditional ML models in detecting the sophisticated phishing patterns.

These feature selection techniques resulted in more efficient computation and less false positive rate due to the reduction of dimensionality of the data and the elimination of redundant attributes. The TabTransformer model was found to have a better contextual learning ability on structured phishing datasets and the GCN model was able to capture the relationships between websites and domain structures.

The framework can be implemented into existing modern cybersecurity infrastructures, and is scalable and reliable for real-world phishing detection systems.

VIII. CONCLUSION

The tactics used in phishing attacks continue to be dynamic and adaptive, making it a significant threat to cybersecurity. The current study suggested an intelligent hybrid phishing detection framework based on feature selection, machine learning models and deep learning models. Experimental results showed that GCN and TabTransformer achieved better classification accuracy than traditional methods. Feature Selection enhanced the efficiency and reduced the complexity of computation. The framework proposed here provides a powerful, scalable and dependable approach to the detection of cutting-edge phishing sites and to improve the security of web surfing.

REFERENCES

[1]. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of KDD*, 2016.
[2]. T. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *International Conference on Learning Representations (ICLR)*, 2017.
[3]. X. Huang et al., "TabTransformer: Tabular Data Modeling Using Contextual Embeddings," *arXiv preprint arXiv:2012.06678*, 2020.
[4]. Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, MIT Press, 2016.
[5]. Verma, R., & Hossain, N., "Semantic Feature Selection for Text with Application to Phishing Email Detection," *IEEE Access*, 2021.
[6]. Bahnsen, A. et al., "Classifying Phishing URLs Using Recurrent Neural Networks," *eCrime Researchers Summit*, 2017.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, March 2026
Impact Factor: 8.2

[7]. S. Haykin, *Neural Networks and Learning Machines*, Pearson Education, 2011.
[8]. Kaggle, "Phishing Website Dataset," Available: https://www.kaggle.com
[9]. UCI Machine Learning Repository, "Phishing Websites Dataset," Available: https://archive.ics.uci.edu
[10]. Singh and P. Sharma, "Machine Learning Approaches for Phishing Detection," *International Journal of Cyber Security*, 2022.