International Journal of Computer Techniques–IJCT Volume 13 Issue 2, March-April - 2026

Open Access and Peer Review Journal ISSN 2394-2231 https://ijctjournal.org/

Intelligent Phishing Detection System Using Machine Learning

T. Murali Gopi Krishna
Computer Science & Engineering
School of Engineering & Technology
Dhanalakshmi Srinivasan University
Trichy, India
muralik4312@gmail.com

T. jaswanth Kumar Reddy
Computer Science & Engineering
School of Engineering & Technology
Dhanalakshmi Srinivasan University
Trichy, India
thanneerujaswanth320@gmail.com

T. Shanmukh Sameer Reddy
Computer Science & Engineering
School of Engineering & Technology
Dhanalakshmi Srinivasan University
Trichy, India
shanmukreddythatiparthi29@gmail.com

Dr.K.Akila
Assistant Professor
Computer Science & Engineering
School of Engineering & Technology
Dhanalakshmi Srinivasan University
Trichy, India
mercygeraldine@gmail.com

Abstract **-** This project presents an advanced phishing detection framework that leverages feature selection along with machine learning and deep learning models such as GCN, TabTransformer, Autoencoder, FNN, and DNN. Using a labeled dataset of legitimate and phishing websites, the system enhances accuracy, generalization, and efficiency through optimal feature selection. Implemented in Python and deployed via a Flask web interface, the framework demonstrates that combining deep learning with feature engineering significantly boosts phishing detection performance, offering a scalable and effective real-world security solution.

Keywords— Phishing Detection, Cybersecurity, Feature Selection, Machine Learning, Deep Learning, Graph Convolutional Network, TabTransformer, Neural Networks, Website Security, Classification Models

I. INTRODUCTION

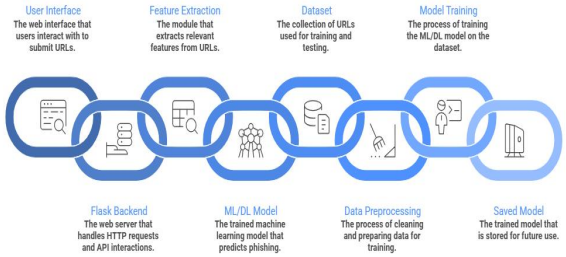
The rise of online services has led to an increase in phishing threats. Phishing is a form of cyberattack where malicious offenders impersonate real companies/organizations to deceive users into giving up their private information (eg. username, password, etc) through either a fraudulent website or email. Since phishing attacks continue to change and evolve as they try to use vulnerabilities for exploitation, they pose a large threat of monetary loss for corporations and people as well.

Older techniques for detecting phishing attacks (such as using a blacklist or a set of rules) have become less viable due to the changing nature of phishing threats in today's world. The reliance upon set rules and previously known attack patterns limits the ability of these conventional methods to detect previously undetected and/or changed phishing websites. Additionally, because of the increased sophistication of the methods used by attackers to hide their phishing websites, it is important that we create intelligent and adaptive solutions that can detect phishing attacks in real time.

In recent times, improvements have been made to enhance machine learning and deep learning-based phishing detection systems by leveraging machine learning models that have previously examined data sets for patterns and physics of histories, as well as deep learning models that capture nonlinear complex features of a website. The performance of different types of model will be dependent upon the features that are passed to them - both in terms of the accuracy of the features and the relation of those features to the task. The addition of many redundant or non-significant features leads to a high dimensional model due to the increased computation expected, which can adversely affect the model when it is applied to a new, unseen (i.e., random) input in a new environment.

To tackle these issues, the research offers an improved phishing detection system that combines optimal feature selection approaches with state-of-the-art machine learning and deep learning models to improve detection accuracy while decreasing the amount of computation required for detecting phishing attempts by selecting the

most relevant features. The proposed framework includes several modeling comparisons between different models such as Graph Convolutional Networks (GCNs), TabTransformer, Autoencoders, Feedforward Neural Networks (FNNs), and Deep Neural Networks (DNNs) and analyses a lab.



This study is aimed at developing a reliable, large-scale, and precise phishing detection mechanism for implementation in real life. The experiments show that utilizing both feature selection and deep learning architecture will improve the capability to detect phishing, and thus are potentially valuable assets in fighting against ever-evolving phishing attacks.eled dataset for both legitimate and phishing websites

II. RELATED WORKS

The area of research into detecting phishing (or other similar strategies) has been widely researched in the area of cybersecurity, with many researchers proposing ways of detecting fraudulent websites and/or messages. In the past, the majority of such research was based on blacklists or heuristics to identify phishing websites; hence, researchers could compare a given URL with a previous, known-to-be-phishing URL or apply some pre-defined "rule" to determine whether the current URL could be considered a phishing site. Although they were successful at finding previously known phishing sites, the vast majority of studies utilizing such methods were unable to detect newly created/phishing URLs due to their high false-negative rates.

Researchers have started looking into using machine learning techniques to solve some of the limitations of previous research. A number of previous studies have applied supervised algorithms (for example, Support Vector Machine, Decision Tree, Naive Bayes and Random Forest Classifiers) to perform phishing detection by using human-engineered characteristics obtained from either the URL or the content of the web site(s) being examined. Although the accuracy and flexibility of the new machine learning approaches performed well relative to traditional methods, their performance was contingent upon the degree of quality and relevance of the machine learning characteristics chosen.

As a result, many researchers focus on finding ways to select the best features from a large data set when building a phishing detection model. Research shows that by reducing the number of irrelevant and redundant features, you can decrease the amount of information contained within a single feature, which can subsequently improve your results' accuracy. There are numerous techniques, including information gain, mutual information, and correlation-based feature selection methods that have been used successfully at reducing dimensionality, improving computational efficiency and increasing generalisation of your model when working with machine learning models on detecting phishing.

Deep Learning models have exploded in popularity through AI's advances to automatically find complex patterns from large amounts of data. Many studies have used different types of neural networks, including Deep Neural Networks, Feedforward Neural Networks, Autoencoders, Convolutional Neural Networks, and Recurrent Neural Networks to detect phishing. Those models have outperformed non-linear feature relationships and have achieved higher detection accuracy compared to traditional machine learning algorithms.

Studies that came out in the last two years have researched new methods of learning from deep networks - including architectures like Graph Convolutional Networks for modeling the structural relationships of website features; and transformer-based architectures for better managing structured, tabular data. The use of these new, advanced methods shows the potential ability of Deep Learning to address complex and evolving phishing attacks. However, the increased computational complexity and low interpretability of many existing deep-learning-based systems when appropriate feature selection processes have not been followed still continues to pose a significant challenge.

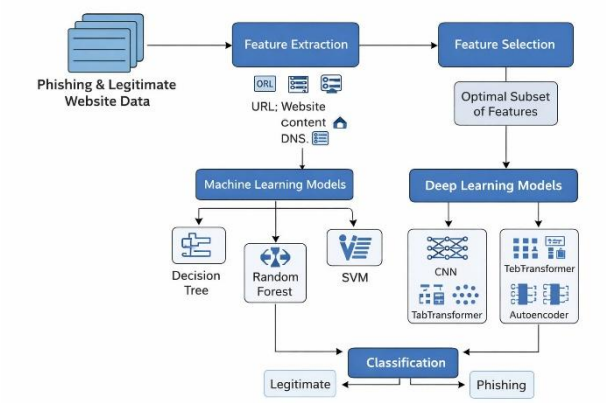
In conclusion, previous studies have shown that machine-learning techniques (e.g., Machine Learning) outperform classic phishing detection techniques. However, there is still a need for an integrated framework that employs both optimal feature selection and advanced deep-learning models to achieve the highest levels of accuracy, scalability, and resilience. Therefore, the proposed study will focus on improving phishing detection by combining feature selection techniques with multiple deep-learning models to provide better overall performance compared to traditional methods.

III. PROPOSED SYSTEM

This suggested system incorporates an intelligent and scalable phishing detection framework which utilizes machine learning, deep learning, and optimal feature selection to accurately detect phishing websites. The overall goal of this system is to improve accuracy and decrease computational complexity and to allow for adaptability of this system as new phishing attacks evolve.

A labeled dataset of both legitimate and phishing sites is created first; every website for which there is training data will have multiple feature dimensions that describe aspects of the URL (the appearance of the URL), the domain name itself and how it behaves. The features would allow model training and other downstream processing to occur normally, as they are the basis for the input to any other process using (or creating).

To solve the issue associated with high-dimensional datasets, this work employs feature-scaling methods to identify the least useless feature. This method will reduce the number of redundant and irrelevant attributes, reducing dimensionality, increasing model capabilities, and providing greater generalizability. Feature scaling is crucial in helping the learning algorithms discriminate against non-useful data that has been collected based on phishing behaviors.



The design of the proposed phishing detection system is shown in the graphic that includes the entire architecture of the system and integrates feature selection with both machine learning and deep learning methodologies to enable accurate identification of phishing websites.The first step in the process is to collect data from both phishing and legitimate websites, which will be the system's input. Data collected includes URL data, web-content data, and domain information data. After raw data is collected, it is processed through feature extraction, leading to the creation of significant features from the URL structure, DNS data, and content data collected from the web page.

After extracting features from a given dataset, the next important step in classification task is to use a feature selection module to identify which features are most relevant to the classification task. This helps to remove redundant and irrelevant attributes, thus minimizing the dimensionality of the feature set and improving the computational efficiency of the classifiers; furthermore, it enhances the capability of a classifier/learner to generalize.Once the optimal feature set has been identified, it will then be sent on to the classification stepIn this classification step, machine learning will have multiple possible model types/algorithms being evaluated in terms of how accurately they can distinguish phishing websites from legitimate/non-phishing websites. Examples of machine learning algorithms used will include abbreviated versions of algorithms such as Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM); among other variants of deep learning algorithms, such as TabTransformer or Autoencoding or any architecture based on deep neural networks that can learn the complex characteristics/patterns and relationships between phishing and legitimate websites.

The last layer is a Classification Layer that divides all the websites into two categories: Legitimate or Phishing. The benefit of using multiple Classification Models is that they can each be compared to one another to analyze performance and will a high level of resiliency against various patterns of Phishing Attacks.Overall this system will take advantage of Feature Extraction, Optimal Feature Selection and Advanced Learning Methods all being done within one system, therefore providing accurate, scalable and effective phishing detection for use in real world applications.

IV. METHODOLOGY

In this research, we aim to create a Phishing Detection Framework that integrates the use of feature selection with machine learning and deep learning techniques in order to develop an effective and smart phishing detection system. We propose a systematic pipeline for implementing the framework, consisting of five stages: data collection, pre-processing, feature extraction, feature selection, training of the model using features selected from previous analysis, and classification of newly identified phishing attacks. Each individual stage has been designed specifically to enhance the accuracy of detection, minimize the computational complexity of using machine learning/deep learning techniques for phishing attack detection, and improve the ability of the

framework to generalize when detecting phishing attacks that have yet to be seen by the system. The flexible, modular structure of this framework will allow for scalability in its application across various types of real-world cybersecurity solutions where phishing detection is required.

Module 1: Data Collection

A labeled dataset containing both legitimate and phishing websites is collected from publicly available repositories.Each data instance includes URL-based, domain-based, and content-based information along with a status label.The dataset serves as the foundational input for all subsequent processing stages.

Module 2: Data Preprocessing

The collected data is cleaned to handle missing values and inconsistencies.Feature values are normalized to maintain uniform scale across attributes.Preprocessing ensures that the data is suitable for machine learning and deep learning model training.

Module 3: Feature Extraction

Relevant features are extracted from website URLs, DNS records, and web content.Extracted features capture structural and behavioral patterns commonly associated with phishing websites.This step transforms raw data into meaningful representations for learning models.

Module 4: Feature Selection

Feature selection techniques are applied to identify the most informative subset of features.Redundant and irrelevant attributes are removed to reduce dimensionality.This module improves model efficiency, accuracy, and generalization performance

Module 5: Model Training

Selected features are used to train multiple machine learning and deep learning models.Models include Graph Convolutional Networks (GCN), TabTransformer, Autoencoder, Feedforward Neural Networks (FNN), and Deep Neural Networks (DNN).Training is conducted under consistent conditions to ensure fair comparison among models.

Module 6: Model Evaluation

Trained models are evaluated using standard performance metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.Comparative analysis is performed to identify the best-performing model.Emphasis is placed on minimizing false positives and false negatives.

Module 7: Classification and Output

The selected model classifies websites into Legitimate or Phishing.Classification results are generated in real time for user interpretation.The system ensures reliable and consistent output across different input samples.

Module 8: Deployment

The final model is deployed using a Flask-based web application.A user-friendly interface developed with HTML and CSS enables real-time interaction.The deployed system supports practical and scalable phishing detection.

V. COMPARISION OF OTHER PAPER

Several research studies have addressed the problem of phishing detection using machine learning and deep learning techniques. However, differences exist in terms of feature selection, model complexity, scalability, and real-time applicability. A comparative analysis of selected existing works with the proposed system is presented below.

Before, Phishing detection focused on rule-based and blacklist rules. It didn't detect many forms of phishing unless they had been seen before. Phishing detection with machine learning improved accuracy, but was often overly reliant on manually crafted features and couldn't adapt to new phishing styles.The available feature selection studies looked mainly at classic machine learning models, which couldn't capture the complex behaviours of phishing.Deep learning methods provided increased recognition of patterns, but many studies have not included specific feature selection leading to increased degression performance.While certain machines were highly accurate, they did not generalise well compared to those that had not been previously seen phishing.

The proposed model employs an optimal integrated solution to feature selection applied to machine Learning as well as Deep Learning models better than what's been done previously.The proposed framework assesses numerous cutting-edge architectures, such as GCN, TabTransformer, Autoencoder, FNN, and DNN as part of a combined process.Feature Selection helps substantially reduce the size (dimensionality) of the data set and thus improves the efficiency of Training and provides better generalization of the results.This approach allows for a comparison and selection of the strongest classification models rather than only using one model.Another significant advantage of this solution is that it allows for real-time phishing detection via a web-based application built using Flask which has been one of the major limitations of past research.Combining all three aspects of quality—scalability, accuracy, and feasibility of deployment—

makes this new solution extremely comprehensive in dealing with current phishing threats.

VI. RESULTS AND DISCUSSION

A labeled dataset with legitimate and phishing websites was used to evaluate the phishing detection framework, comprised of the same dataset and a common set of features for evaluation. Both traditional machine learning and deep learning models were trained and tested on the same dataset and features to provide an unbiased comparison. The performance of all classifiers was evaluated using the following evaluation metrics: accuracy, precision, recall, F1-score, and AUC-ROC.

Results from this research demonstrate that integrating feature selection in each of the classifiers substantially improves the performance of the model on all classifiers evaluated. Models that utilized feature selection were able to more quickly converge and reduce their computational complexity while also improving their classification accuracy by eliminating irrelevant/redundant features.

Traditional machine learning methods were outperformed by deep learning methods. For example, the use of DNNs and channelized FNNs was very effective because they can model non-linear relationships in the feature space very well, leading these models to produce very high-quality solutions. Nevertheless, the TabTransformer was also observed to be effective with regards to structured tabular data, learning feature interactions through the application of attention methods. The use of Autoencoders also enhanced anomaly detection when used to identify problematic behavior typically associated with phishing activities.

The results of the study provide several insights into how to effectively detect phishing. One major conclusion is that feature selection is paramount to improving detection rates through reducing noise and dimensionality. All models created using selected features outperformed models created using the entire feature set, which enables us to conclude that optimizing features enhances the application of cybersecurity.The second conclusion from the study is that deep learning algorithms outperformed other types of algorithms (i.e., traditional machine learning) to detect phishing. The increased performance of deep learning can be attributed to its ability to learn complex patterns and interactions between data that cannot be learned as easily with shallow models. However, deep learning algorithms also have an increased computational cost and interpretation difficulty when using all features. The proposed framework for detecting phishing mitigates this drawback by incorporating feature selection before the training phase of both deep learning and traditional machine learning models.

Third, utilizing multiple models coordinated with a unified framework permits comparative analysis as well as robustness. The ability to evaluate many different architectures and then choose the best performing model based on several established performance metrics provides more reliability and adaptability against changing phishing techniques than simply using one classifier.Lastly, the usability of the selected classifier through a web-based application demonstrates the potential usefulness of the entire proposed methodology. This capability allows for real-time phishing detection, which makes this method applicable to real-time situations, including browser-based security products and enterprise cybersecurity solutions.

VII. CONCLUSION AND FUTURE SCOPE

The rapidly evolving terror of phishing attacks presents a sobering challenge to protecting cybersecurity as they become increasingly creative in exploiting user trust. In the present study, an improved phishing detection framework was proposed that incorporates optimized feature selection combined with both machine and deep learning models to effectively distinguish between phishing and legitimate websites.

The system being proposed has been developed in such a way that there is a step-by-step process for taking the data through the process of being prepared, breaking the data down into its individual parts, selecting the right parts of that data, creating and training a model, and finally, using that model to predict new data. Selecting the best features allows the system to reduce the amount of data involved (dimensionality), improve the speed of computation, and provide better generalization of the model. When evaluating different models (e.g., Graph Convolutional Networks, TabTransformer, Autoencoder, Feedforward Neural Networks, and Deep Neural Networks), it has been shown that deep learning methods using optimized feature sets are superior to traditional techniques.

The experimental results confirm that the proposed framework provides improved detection accuracy, lower false positive rates and better robustness against previously unseen phishing attacks. Additionally, the deployment of the best model on a web interface demonstrates the practicability of the system in real-world settings. Overall, this research provides a scalable and effective method for detecting phishing and addresses major shortcomings of current approaches to enhance defenses against contemporary phishing attacks.

References

1] F. Salahdine, Z. E. Mrabet, and N. Kaabouch, "Phishing attacks detection a machine learning-based approach," in Proc. IEEE 12th Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON), New York, NY,USA, Dec. 2021, pp. 250-255.

[2] M. Baykara and Z. Z. Gürel, "Detection of phishing attacks," in Proc.6th Int. Symp. Digit. Forensic Secur. (ISDFS), Antalya, Turkey, Mar. 2018,pp. 1-5.

[3] K. L. Chiew, K. S. C. Yong, and C. L. Tan, "A survey of phishing attacks: Their types, vectors and technical approaches," Expert Syst. Appl., vol. 106, pp. 1-20, Sep. 2018.

[4] F. Yahya, "Detection of phishing websites using machine learning approaches," in Proc. Int. Conf. Data Sci. Appl. (ICoDSA), Bandung, Indonesia, 2021, pp. 40-47

5] A. Alswailem, B. Alabdullah, N. Alrumayh, and A. Alsedrani," Detecting phishing websites using machine learning," in Pr oc. 2nd Int. Conf. Comput. Appl. Inf. Secur. (ICCAIS), Riyadh, Saudi Arabia, May 2019, pp. 1- 6.

[6] H. Zuhair, A. Selamat, and M. Salleh," Feature selection for phishing detection: A review research," Int. J. Intell. Syst. Technol. Appl., vol. 15,no. 2, p. 147, 2016.VOLUME 13, 2025 33319

[7] Mendeley Data. (Sep. 2020). Phishing Websites Dataset. [Online]. Available: https://data.mendeley.com/datasets/72ptz43s9v/1 [8] Mendeley Data. (2021). Web Page Phishing Detection Dataset. [Online].Available: https://data.mendeley.com/datasets/c2gw7fy2j4/3

[9] P. Chinnasamy, N. Kumaresan, R. Selvaraj, S. Dhanasekaran,K. Ramprathap, and S. Boddu, An efficient phishing attack d etection using machine learning algorithms," in Proc. Int. Conf. Advancements Smart, Secure Intell. Comput. (ASSIC