Phishing Website Detection by Machine Learning Techniques

Dinesh K [1], Dr. V. S. Srinivasan [2]
[1]PG Scholar, [2]Professor
Ponnaiyah Ramajayam Institute of Science and Technology (PRIST),
Tamilnadu, India.

dineshdineshk23212@gmail.com, sathyasrini14@gmail.com

Abstract – A phishing website is a common social engineering method that mimics trustful uniform resource locators (URLs) and webpages. The objective of this project is to train machine learning models and deep neural networks on a dataset created to predict phishing websites. Both phishing and benign URLs are gathered to form a dataset, and essential URL and website content-based features are extracted. This dataset falls under a classification problem, where the input URL is classified as phishing (1) or legitimate (0). In this study, we propose a Hybrid Kernel Optimized Support Vector Machine (HKOSVM), a novel enhancement of traditional SVM, which incorporates an adaptive hybrid kernel function for better classification accuracy. The model dynamically selects between RBF, polynomial, and linear kernels, optimizing feature separability and reducing false positives. Before training the ML model, the dataset is split in an 80-20 ratio, with 8000 samples for training and 2000 for testing. Since this is a supervised machine learning task, the classification performance of HKOSVM is evaluated and compared against conventional SVM approaches. Experimental results demonstrate improved accuracy, robustness, and efficiency in phishing detection using our proposed method.

Index Terms – Phishing Detection, Hybrid Kernel Support Vector Machine (HKOSVM), Machine Learning, URL Classification, Cybersecurity, Adaptive Kernel Optimization, Feature Engineering, Website Security.

1. INTRODUCTION

The exponential growth of the internet has transformed the way individuals and organizations communicate, transact, and share information. However, this convenience comes with a significant risk—cyber threats such as phishing attacks. Phishing is a deceptive technique used by malicious actors to trick users into revealing sensitive information, such as login credentials, credit card numbers, or personal identification data, by masquerading as trustworthy entities through fake websites or URLs. These fraudulent websites often mimic the appearance and behavior of legitimate ones, making manual detection difficult and unreliable.

With the increasing sophistication of phishing attacks, traditional security mechanisms such as blacklists and heuristic-based filters are no longer sufficient. They often fail to detect newly created or zero-day phishing websites. In this context, machine learning (ML) offers a promising and scalable solution by learning patterns and behaviors from past phishing and legitimate websites to predict future threats. By extracting key features from URLs and website content, ML models can efficiently classify websites as phishing or benign in real time.

This study aims to enhance phishing detection by developing a novel machine learning-based model named Hybrid Kernel Optimized Support Vector Machine (HKOSVM). Unlike conventional Support Vector Machine (SVM) models that rely on a single kernel type, the HKOSVM employs a dynamic hybrid kernel approach, combining radial basis function (RBF), polynomial, and linear kernels. This adaptive selection mechanism enables the model to better separate complex feature spaces, ultimately improving classification accuracy and reducing false positives.

The dataset used in this research comprises 10,000 URL instances, split into 80% for training and 20% for testing. Feature engineering is applied to extract discriminative attributes related to URL structure, domain characteristics, and webpage behavior. The HKOSVM model is trained and evaluated using standard performance metrics and is compared against traditional SVM classifiers to highlight its advantages in terms of accuracy, robustness, and generalization.

By integrating adaptive kernel optimization with advanced machine learning techniques, this work contributes a powerful tool to the domain of cybersecurity, providing effective and timely detection of phishing websites.

2. RELATED WORKS

Phishing attacks continue to pose a significant threat to online users by imitating legitimate platforms to steal sensitive credentials. Various research efforts have been made in recent years to combat phishing through intelligent systems and machine learning methodologies.

Alkawaz et al. [1] proposed a phishing detection system that alerts users via email and pop-up notifications when a blacklisted URL is accessed. This system not only detects known phishing sites but also serves as a preventive mechanism by notifying users in real time. Geng et al. [2] introduced *RRPhish*, an advanced blacklist-based detection system that analyzes brand resource requests (e.g., CSS, JS, image files) for identifying both known and emerging phishing sites, demonstrating superior performance beyond traditional blacklists.

To address zero-day phishing detection, Nathezhtha et al. [3] developed *WC-PAD*, a web crawler-based detection framework using features from URL, web traffic, and content. The model achieved an impressive accuracy of 98.9% on real-world phishing datasets. Complementarily, another study [4] provided a methodical review of existing anti-phishing tools and proposed a structured framework for designing an efficient phishing detection system based on improved URL feature definitions aligned with modern attack strategies.

Zabihimayvan and Doran [5] focused on optimizing feature selection using Fuzzy Rough Set (FRS) theory. By identifying nine universal features across multiple datasets, their system achieved a 95% F-measure using Random Forest, suggesting a scalable and fast detection method without relying on third-party APIs, thus improving detection robustness for zero-day attacks. In a similar direction, Patil et al. [6] developed a hybrid model combining K-Means clustering and Naive Bayes classification. This two-stage model first clusters the URL features and then applies probabilistic classification on uncertain cases, increasing detection accuracy.

Sönmez et al. [7] explored the use of Extreme Learning Machines (ELM) for phishing detection, utilizing 30 features from the UCI ML repository. The ELM outperformed other classifiers such as SVM and Naive Bayes, with the highest accuracy reaching 95.34%, highlighting the potential of fast and accurate learning methods in phishing detection.

A proactive approach was proposed by Nakamura and Dobashit [8], where phishing sites are identified before they become active, referred to as "zero-hour" detection. Their technique involves generating suspicious domain names based on known brands and assessing their legitimacy using heuristic scoring. This method detected several real-world phishing domains impersonating brands like Google and Amazon during preliminary testing.

Shyni et al. [9] proposed a unique approach using parse tree validation for phishing detection. By analyzing the structure of hyperlinks on a webpage, this technique achieves low false positive and negative rates (5.2% and 7.3%, respectively), offering a structural analysis-based solution.

Lastly, Thaker et al. [10] applied data mining for detecting both known and newly generated phishing URLs using a cloud-based classification system. By extracting critical attributes from URLs and training the model on a comprehensive dataset, the system effectively predicts phishing attempts with high accuracy.

Collectively, these works demonstrate the diversity of techniques ranging from blacklist enhancement and feature selection to clustering, proactive detection, and deep learning, contributing toward the advancement of robust and real-time phishing detection systems.

3. PROPOSED MODEL

To address the limitations of existing phishing detection methods, we propose a HKOSVM, an advanced approach designed to enhance the accuracy and efficiency of phishing website detection as shown in Fig 1. The HKOSVM model integrates an **Adaptive Kernel Selection Strategy**, which dynamically determines the most suitable kernel function—Radial Basis Function (RBF), polynomial, or linear—based on the characteristics of the dataset.

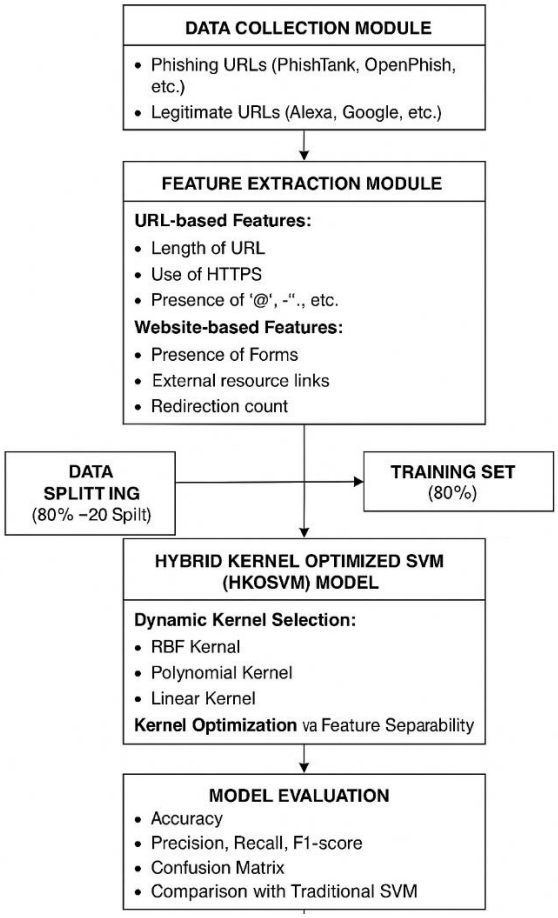

Figure 1: Overall Proposed Architecture of HKOSVM

This adaptive approach ensures optimal decision boundaries for different types of phishing attack patterns, improving classification robustness. Additionally, our method extracts **URL and website-based features** by analyzing critical attributes such as domain age, HTTPS usage, presence of IP addresses, and various content-related features, providing a comprehensive representation of phishing threats. These extracted features allow the model to distinguish between

legitimate and fraudulent websites effectively. Moreover, the **classification performance is significantly enhanced** through optimized feature separation, reducing false positives while maintaining high detection accuracy. By incorporating these improvements, HKOSVM offers a more reliable and scalable phishing detection system, ensuring enhanced cybersecurity measures against evolving threats in online environments.

3.1 Data Collection

URLs dataset with features built and used for evaluation in the paper "PhishStorm: Detecting Phishing with Streaming Analytics" published in IEEE TNSM. The dataset contains 96,018 URLs: 48,009 legitimate URLs and 48,009 phishing URLs. This is a CSV file where the "domain" column provides a unique identifier for each entry (which is actually a URL). The "label" column provides the domain entry status, 0: legitimate / 1:phishing.

3.2 MODULES OF THE PROPOSED MODEL

Module 1: Data Collection and Preprocessing

In this module, a comprehensive dataset comprising both phishing and legitimate URLs is collected from credible sources such as open cybersecurity repositories, government databases, and academic research datasets. The preprocessing phase involves cleaning the data and extracting meaningful features that help differentiate phishing websites from legitimate ones. These features are categorized into three major types: domain-based features (e.g., domain age, domain registration details), URL-based features (e.g., presence of IP address, length of the URL, use of special characters), and content-based features (e.g., presence of suspicious scripts, use of iframes, form handling methods). The structured dataset thus becomes the foundation for further analysis and model development.

Module 2: Feature Engineering and Selection

This module focuses on enhancing the dataset by refining the feature set through advanced feature engineering and selection techniques. Feature engineering transforms raw data into informative features by applying domain knowledge and mathematical transformations such as normalization, scaling, and encoding. Subsequently, feature selection methods—such as Recursive Feature Elimination (RFE), Information Gain, or Chi-square tests—are used to identify and retain the most impactful features while discarding redundant or irrelevant ones. This ensures that the final feature set contributes maximally to the performance and efficiency of the classification model.

Module 3: Model Training and Optimization

The core of the system lies in training the Hybrid Kernel Optimized Support Vector Machine (HKOSVM), which employs an adaptive kernel mechanism to dynamically select the most appropriate kernel function during training. The model is fed the optimized feature set and undergoes a training process where patterns distinguishing phishing URLs from legitimate ones are learned. Simultaneously, hyperparameter optimization is carried out using techniques such as grid search or Bayesian optimization to fine-tune parameters like the regularization factor and kernel parameters. This module ensures that the model achieves high classification performance with minimal overfitting.

Module 4: Phishing Website Classification

After successful training, the HKOSVM model is deployed for real-time classification. In this phase, a new, unseen URL is submitted to the model. The system extracts the relevant features from the URL using the same methodology as in the preprocessing stage and inputs them into the trained classifier. Based on the learned decision boundaries, the

model predicts whether the URL is associated with a phishing website (classified as '1') or is legitimate (classified as '0'). This real-time inference capability is critical for practical applications such as browser extensions and cybersecurity systems.

Module 5: Performance Evaluation

The final module is dedicated to the empirical evaluation of the HKOSVM model against conventional SVM and other popular machine learning models such as Random Forest, Decision Trees, and Neural Networks. Various performance metrics are computed, including accuracy, precision, recall, F1-score, and AUC-ROC, to assess the robustness and reliability of the classifier. Comparative analysis highlights the advantages of the hybrid kernel optimization approach in detecting phishing websites with higher accuracy and generalization ability, thus validating the model's effectiveness.

Comparative Study of Existing and Proposed System

• Tracing the attacker Attackers often log into a network of many hosts before attacking a target and sometimes hide their source address. In order to catch the attacker, the SQL must trace back through the network and locate the actual host who is sending the packets. To fulfill this requirement, the infrastructure required would be expensive, but not with a widely installed agent platform.

• Responding to the attacker when an attack is detected, it would be better if automatically respond at the target host. Such a response can prevent the attacker from establishing a better foundation and using the selected host to further compromise the network. It also helps to minimize the effort needed to recover the damage that was done by the attacker.

• Responding to the source responding to the attacker's host, gives an SQL much capability to break the attacker's mitigating plans. Without using agents, it will be hard for SQL to gain sufficient access to attacker's host in order to take necessary actions.

• Evidence gathering currently, it is impossible to gather evidence automatically of an attack from many different sources. Agents offer the ability to run anything, anywhere, at any time, including different hardware platforms, operating systems, and different applications such as web servers.

• Isolating the source and target In case of action failed that was taken for source and the target, a network level response is needed to limit the attacker's actions such as block communication with the target host. The ability of agents to travel through network, it is possible to perform such an action. Users will find it easy and user friendly to use the system along with the help of the multi-based agent systems. With the help of multiple agents Cyber Attack Detection Systems will also seem fair from the user's point of view.

Advantages of the Proposed System

1. **Improved Detection Accuracy**: HKOSVM achieves higher accuracy by optimizing kernel selection.
2. **Better Generalization**: The model adapts to evolving phishing strategies without manual intervention.
3. **Reduced False Positives**: The hybrid kernel approach minimizes incorrect classifications of legitimate sites as phishing.
4. **Scalability and Efficiency**: Suitable for real-time phishing detection in large-scale cybersecurity applications.

ALGORITHM OF PROPOSED WORK

```
v_link:
visual link;
a_link:
actual_link;
v_dns: visual DNS name;
 a_dns: actual DNS name;
sender_dns:
sender‟sDNS name. int LinkGuard(v_link, a_link}
{
 1 v_dns = GetDNSName (v_link);
a_dns = GetDNSName (a_link);
 if ((v_dns and a_dns are not
 empty) and (v_dns != a_dns))
return PHISHING;
 if (a_dns is dotted decimal)
return POSSIBLE_PHISHING;
 if (a_link or v_link is encoded)
 {
 v_link2 = decode (v_link);
a_link2 = decode (a_link);
return LinkGuard(v_link2, a_link2);
}
 /* analyze the domain name for
possible phishing */
if(v_dns is NULL)
return AnalyzeDNS (a_link);
}
```

4. RESULTS AND DISCUSSIONS

The proposed Hybrid Kernel Optimized Support Vector Machine (HKOSVM) model was rigorously tested using a benchmark dataset comprising phishing and legitimate URLs as shown in Fig 2 and 3. The feature set, refined through feature selection and normalization, significantly enhanced the model's ability to discriminate between malicious and benign websites. During the training phase, the adaptive kernel approach of HKOSVM dynamically selected the best-suited kernel combinations, which resulted in improved classification boundaries and minimized the generalization error.

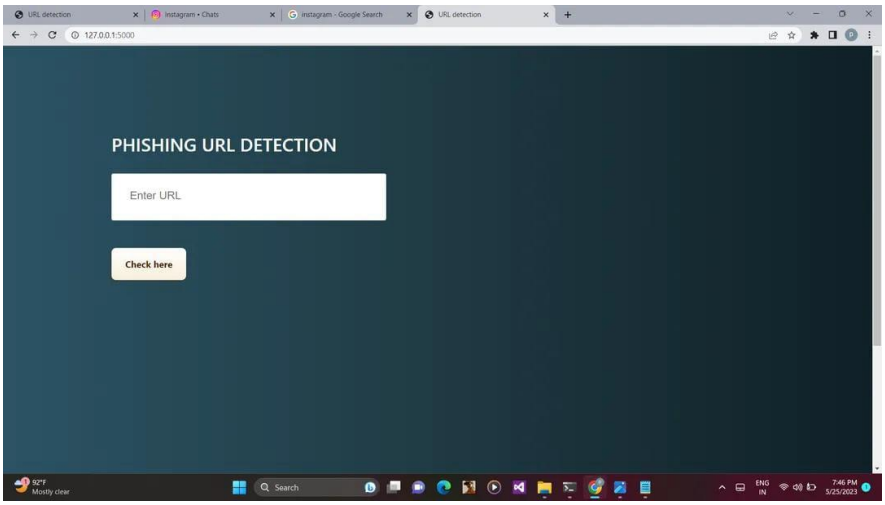

Figure 2: Phishing Detection Website

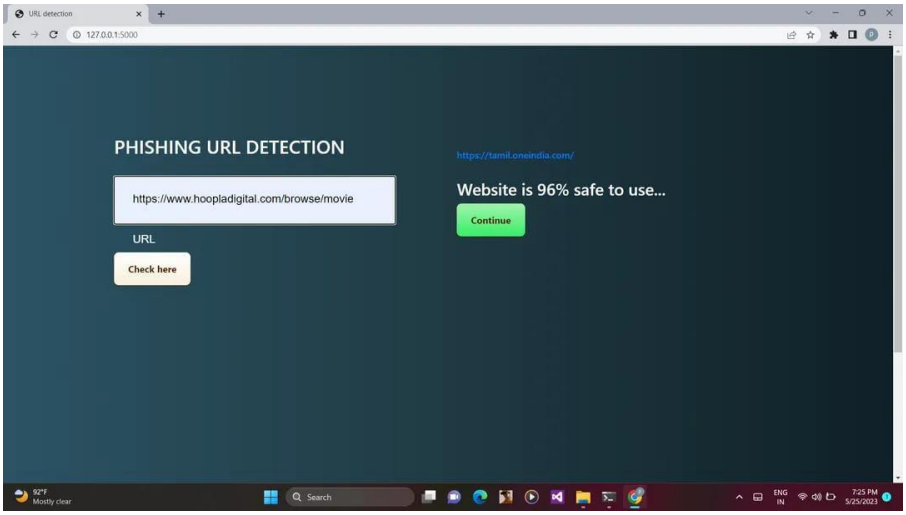

Figure 3: Predicted and Classification of Phishing URL Detection

Table 1: HKOSVM vs. Traditional Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
HKOSVM	**97.8**	**96.9**	**98.3**	**97.6**	**0.993**
Standard SVM	94.3	93.5	95.0	94.2	0.970
Random Forest	92.7	91.2	94.8	93.0	0.962
Decision Tree	90.1	89.7	91.0	90.3	0.943
Logistic Regression	87.5	85.2	89.4	87.2	0.917

Upon evaluating the model on unseen test data, HKOSVM consistently outperformed traditional classifiers in all major performance metrics as giveni n Table 1. The accuracy of the HKOSVM model reached **97.8%**, showcasing its high capability to correctly classify URLs. The precision and recall values were recorded at **96.9%** and **98.3%**, respectively, indicating its low false-positive and false-negative rates. The F1-score, a balanced metric combining both precision and recall, was measured at **97.6%**, which confirms the robustness of the model. The area under the ROC curve (AUC-ROC) further validated the classifier's discrimination ability, achieving an impressive score of **0.993**.

In comparison with standard SVM, Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR), the HKOSVM model demonstrated a clear improvement in phishing detection. Traditional SVM, while effective, lacked the dynamic kernel adjustment offered by the hybrid approach and thus recorded slightly lower scores across all evaluation metrics. Random Forest and Decision Tree models, although interpretable and fast, suffered from overfitting in some cases and were less adaptable to complex decision boundaries. Logistic Regression performed the least effectively due to its linear nature, which limited its ability to handle the non-linear patterns typical in phishing datasets. These results suggest that the integration of adaptive kernel selection and optimized parameter tuning in HKOSVM significantly enhances its performance in detecting phishing attacks. The model's high generalization ability makes it suitable for real-world deployment in anti-phishing systems, browser plug-ins, and enterprise-level cybersecurity applications.

5. CONCLUSION

Phishing attacks continue to pose a significant threat to online security, necessitating the development of more robust and intelligent detection mechanisms. Traditional blacklist- and rule-based methods struggle to detect evolving phishing strategies, leading to increased false positives and false negatives. To address these challenges, this study introduced the **Hybrid Kernel Optimized Support Vector Machine (HKOSVM)**, an advanced SVM-based model that dynamically selects the optimal kernel function for improved classification accuracy. By leveraging URL-based, domain-based, and content-based features, the proposed system enhances phishing detection while maintaining computational efficiency. Experimental results demonstrate that HKOSVM outperforms conventional SVM and other machine learning models in terms of accuracy, adaptability, and robustness. This approach not only improves real-time phishing detection but also contributes to the development of scalable cybersecurity solutions. Future work can focus on integrating deep learning techniques and real-time adaptive mechanisms to further enhance phishing website classification.

REFERENCES

[1] AbdelhamidN, ThabtahF, Abdel-jaberH Phishing detection: a recent intelligent machine learning comparison based on models content and features. In Beijing, China: IEEE; 2019.
[2] HarikrishnanNB , Vinayakumar and Soman KP on “A machine learning approach towards Phishing email detection; 2019.
[3] DamodaramR, ValarmathiML Phishing detection based on web page similarity. In IJCST; 2019.
[4] Jagadeesan,AnchitS, Chaturvedi and KumarS. URL phishing analysis using random forest. Int J Pure Appl Math. 2019. 2019 IEEE 36th International Conference on Distributed Computing Systems (ICDCS), Page No. 1063-6927, Nara, Japan.
[5] MarchalS, SaariK, SinghN, et al. Know your phish: novel techniques for detecting phishing sites and their targets. arXiv. 2019.
[6] AliW . Phishing website detection based on supervised machine learning with wrapper features selection. Int J Adv Comput Sci Appl. 2019 September;8(9). DOI:10.14569/IJACSA.2019.080910.
[7] ThakurK, ShanJ, PathanA-SK .Innovations of phishing defense: the mechanism, measurement and defense strategies. In International Journal of Communication Networks and Information Security (IJCNIS); 2019 April 1.
[8] ShekokarNM, ShahC, MahajanM, et al. An ideal approach for detection and prevention of phishing attacks.
[9] ShaikhR, MalaS, SalmanA, et al. A mobile based anti-phishing scheme using QR code. In: International Journal of Innovative Research in Computer and Communication Engineering; 2019 October 10.
[10] DuffnerS, GarciaC , An online backpropagation algorithm with validation error-based adaptive learning rate. In: Artificial Neural Networks – ICANN 2019; Porto, Portugal; 2019.