# An Interactive Machine Learning Apporach with H-Svm for Url Phishing Detection in Real Time Analysis

**[1]Srinivasa Rao Dammavalam, [2]ThalariVigneshwar**

[1] Associate Professor, Department of IT,VNR VignanaJyothi Institute of Engineering andTechnology,Hyderabad, India-500090

[2]M.Tech Student, CNIS, VNR VignanaJyothi Institute of Engineering andTechnology,Hyderabad, India-500090

**Abstract:**

The importance on the how to tackle and subside the URL-Phishing problems have led to remarkable changes in the aspect of the Web technologies. While the importance and different learning algorithms multiple aspects of phishing models have been improvised with ML or DL methods to suffice the impact on the real time analysis.

The design aspect of proposed model, determines the combination on the Hybrid SVM method with other algorithm based on deep learning techniques on the aspect of URL-phishing detection as observed to be accuracy of 97.1%. The importance of the SVM method describes on the different feature extraction process of the dataset chosen from UCI website. To realize the effective importance on the predictive analysis with SVM based results tends to be adaptable with the choices to improvise the cybercriminals as the SVM provides linear patterns. To effectively solve the problem of the recognition on linear patterns we implicate a Hybrid aspect of the design on SVM with deep learning model to monitor and realize the correct predictive analysis for in-on time analysis to reduce the possible chances of attack.

***Keywords: URL, SSL, phishing attacks, SVM, dataset.***

## I Introduction

Attacks involving phishing provide an enduring or substantial danger to cybersecurity by persistently capitalising on individuals susceptibilities to compromising confidential data, identification, and monetary assets. These assaults make use of a range of social engineering strategies in order to manipulate users into revealing their personal and private information. A prevalent and efficacious strategy used by individuals engaging in phishing activities involves the creation of false login URLs that closely resemble authentic websites, with the intention of enticing users to disclose their login credentials. The effect of phishing assaults has been magnified in recent years due to the widespread availability of internet-based services and the growing dependence on digital platforms. The principal objective of phishing is to manipulate users, inducing them to inadvertently disclose their login credentials. These authentication details are then abused by cybercriminals for the purpose of financial gain or unauthorised access to confidential information. The utilisation of authentication URLs in phishing scams is a matter of special concern owing to its direct correlation with consumers' confidence and dependence on genuine online services. These attacks exploit users' reliance on familiar login pages, causing them to disregard apparent inconsistencies that may signal a deceptive website. The efficacy of conventional phishing detection systems is often hindered by the rapid evolution of attacker tactics. Consequently, it is vital that researchers create sophisticated strategies that can accurately distinguish phishing URLs in real-world situations.

The primary objective of this study is to tackle the difficulties presented by phishing attempts that use misleading login URLs. The objective of this project is to improve the accuracy of identifying fake URLs by developing and evaluating a complete approach for phishing URL identification. This will be achieved through the analysis of real-case situations. By using machine learning methods, doing online content analysis, and considering URL characteristics, this approach aims at offering a strong defence mechanism against the increasing menace of phishing attempts. In the

following sections, we will explore the particular methodology used, the datasets utilised for both training and assessment purposes, the results acquired, and the implications derived from the findings. This investigation makes a valuable contribution to the overall endeavours of improving web safety and safeguarding customers from phishing assaults that use fake login URLs by upgrading the state of technology used for phishing detection.

**Problem statement:**

The enduring and dynamic characteristics of phishing assaults provide a significant obstacle to the field of cybersecurity, specifically in relation to the implementation of misleading login URLs that imitate authentic websites in order to deceive users into revealing confidential login information. Conventional strategies for combating phishing encounter difficulties due to the continuous adaptation of attackers, who abuse users' reliance on recognisable authentication interactions and employ sophisticated methods to deceive users. The primary objective of this study is to design a resilient and efficient technique that precisely addresses the challenge posed by these intricate phishing URLs. This methodology aims to include advanced technologies such as machine learning, online content analysis, and evaluation of URL characteristics in order to address the constraints of current approaches and effectively combat the dynamic strategies employed by cybercriminals. The objective of the approach is to effectively identify fraudulent login pages by taking into account the dynamic and varied characteristics of real-life situations. It aims to achieve a balance between accuracy and minimising false positives while also being flexible enough to address new threats without the need for regular human updates.

The main objective of this study is the development of detection methods that possess the ability to recognise novel and intricate phishing URL variations, with a specific emphasis on those that use misleading login pages. The study endeavours to deliver successful responses by using a comprehensive methodology that integrates machine learning, online analysis, and developments in cybersecurity. This initiative is expected to greatly enhance the security of online users, organisations, and digital ecosystems in the face of the harmful menace of phishing assaults, eventually strengthening defences against these deceptive cyber threats that exploit login URLs.

**Ii Literature Survey**

In order to achieve effective phishing operations, attackers use many strategies, including the dissemination of emails that possess an appearance of trustworthiness. These emails may have a multitude of alterations, such as the inclusion of the recipient's name inside the subject line or the modification of URLs within the body of the email. These strategies are used to circumvent filtering mechanisms and create challenges for information system teams in their efforts to effectively block all emails, even in cases when they possess knowledge of an ongoing assault. Insufficient attention is given to the practise of categorising emails into campaigns, which aims to enhance the assistance provided to personnel involved in the identification and mitigation of phishing attacks via reported instances of phishing. This study in [1] investigates the viability of using clustering algorithms to categorise emails into campaigns that may be seen as comparable by IT personnel. Initially, the Meanshift and DBSCAN algorithms were used using a total of seven distinct feature sets. Subsequently, an assessment was conducted on the solutions using the Silhouette coefficient and homogeneity score. The results revealed that Mean Shift exhibits superior performance compared to DBSCAN when considering variables related to email origin and URLs. Subsequently, a user survey is conducted in order to authenticate the efficacy of our clustering method, revealing that clustering exhibits promise as a viable strategy for campaign identification.

The primary concern in cybersecurity threats is the presence of malevolent Uniform Resource Locators (URLs). Cyber assailants disseminate malevolent URLs in order to execute various forms of assaults, including phishing and malware, so ensnaring unwary individuals in fraudulent schemes that culminate in financial detriment and the compromise of sensitive data. The proliferation of Quick Response (QR) codes containing malicious URLs has emerged as a

pressing concern, representing an ongoing security challenge. Most current QR link detection scanner programmes mostly rely on the blacklist approach to identify potentially harmful URLs. However, this strategy is not the most effective for identifying newly created websites. In recent times, there has been a surge in the use of machine learning techniques for the purpose of augmenting the identification of malevolent URLs. Nevertheless, these methodologies are fully reliant on data and need a substantial and up-to-date dataset for training in order to develop a proficient detection approach. This study presents QsecR, a QR code scanner that prioritises security and privacy, based on a framework for detecting dangerous URLs. QsecR is an Android application that functions as a QR code scanner. It operates by using a predetermined static feature categorization system, which has 39 classes of blacklist, lexical, host-based, and content-based characteristics. A collection of 4000 URLs, randomly selected from real-world sources such as URLhaus and PhishTank, was compiled for the dataset. The security and privacy aspects of the QsecR are assessed by many QR code readers. The testing findings demonstrate that QsecR exhibits superior performance compared to other methods, attaining a detection accuracy of 93.50% and a precision value of 93.80%. These in [2] results indicate a substantial improvement over existing secure QR code scanners. QsecR stands out as a very privacy-conscious programme, characterised by its minimal permission requirements.

Presently, a multitude of cybercriminal activities are orchestrated over the internet. Therefore, the primary objective of this research in [3] is to examine phishing attempts. Despite its initial emergence in 1996, phishing has evolved into a very consequential and perilous kind of cybercrime inside the realm of the internet. Phishing employs email manipulation as its fundamental tool for deceptive communications, along with counterfeit websites, in order to illicitly acquire the necessary information from targeted individuals. Various research have been conducted to explore the precautionary measures, detection techniques, and awareness around phishing assaults.

However, a comprehensive and effective approach to effectively counteract these attacks is presently lacking. Hence, machine learning assumes a crucial role in the mitigation of cybercrimes pertaining to phishing attempts. The work is centred on a dataset of phishing URLs obtained from a well-known repository of datasets. This dataset includes properties of both phishing and genuine URLs, which were gathered from over 11,000 website datasets and represented in vector format. Following the preprocessing stage, many machine learning techniques have been used and specifically tailored to mitigate the risk of phishing URLs, hence enhancing user safety. This research employs various machine learning models, including decision tree (DT), linear regression (LR), random forest (RF), naive Bayes (NB), gradient boosting classifier (GBM), K-neighbors classifier (KNN), support vector classifier (SVC), and a novel hybrid LSD model. The hybrid LSD model combines logistic regression (LR), support vector machine (SVC), and decision tree (DT) using soft and hard voting techniques. The objective of this study in [4] is to effectively and accurately mitigate phishing attacks. The suggested LSD model uses the canopy feature selection approach in conjunction with cross-fold validation and Grid Search HyperparameterOptimisation techniques. In addition, many evaluation metrics were used to assess the suggested technique, including precision, accuracy, recall, F1-score, and specificity. These characteristics were employed to demonstrate the impact and effectiveness of the models. The comparison studies provide findings that indicate the superiority of the suggested strategy over the other models, resulting in the attainment of optimal outcomes.

Phishing is a significant security problem that has profound implications for both people and the companies that are targeted. Despite the enduring presence of this danger, it continues to exhibit significant levels of activity and efficacy. Indeed, the strategies used by assailants have undergone continual evolution throughout the years, with the aim of enhancing the persuasiveness and efficacy of their assaults. In the present setting, the identification of phishing has paramount significance. The existing body of literature presents a wide range of techniques that address this problem, specifically focusing on the identification and detection of phishing websites.

This study in [5] offers an extensive and inclusive examination of the current state of knowledge in this particular topic via an analysis of the primary obstacles and discoveries. The focus of the debate is on three significant classifications of detection methods, notably list-based, similarity-based, and machine learning-based techniques. In this study, we provide a comprehensive overview of the detection techniques provided in the existing literature for each category. Additionally, we examine the datasets that have been used to test these approaches. Furthermore, we identify and address certain research gaps that need further investigation and exploration.

Social engineering assaults are predicated upon the exploitation of human fallibility and decision-making processes. The semantic assault is a kind of social engineering attack that involves the use of deceptive tactics, either in behaviour or appearance, to manipulate others. For instance, an attacker may develop a malicious website that closely resembles and functions similarly to a real website. The prevalent forms of social semantic assaults include phishing, spamming, defacement, and malware attacks. In this study, we explore in [6] the viability of constructing social semantic assault detection models based on URLs, using character-aware language models. We have constructed three distinct models, namely the long short-term memory (LSTM)-based detection model, the convolutional neural network (CNN)-based detection model, and the CharacterBERT-based detection model. We conducted a performance evaluation of many models across different assault scenarios. The characterBERT-based detection model achieved a high detection accuracy of 99.65% with the use of a 5-fold cross-validation approach for overall assessment. In terms of individual class performance, the CharacterBERT model demonstrated superior performance compared to the other two models in recognising social semantic assaults. It achieved the highest accuracy of 99.90% specifically in identifying defacement attacks.

Phishing assaults have shown a notable increase in prevalence within the realm of cybercrime in recent times. Social engineering assaults include the manipulation of individuals via the dissemination of fraudulent messages through social media platforms or email communication channels. Phishing attacks are designed to illicitly get users' personal information or surreptitiously install dangerous software. The difficulty in detecting phishing attempts arises from the ability of attackers to craft deceptive messages that closely resemble real communications, so deceiving the recipient. The present communication potentially includes a URL associated with phishing, so highlighting the susceptibility of even proficient individuals to falling prey to such fraudulent activities. The provided URL directs them to a fraudulent website that illicitly acquires sensitive data, including login credentials, payment details, and other pertinent information. Researchers and engineers collaborate to develop techniques for detecting phishing assaults autonomously, therefore eliminating the need on human expertise. Even though numerous studies describe HTML and URL-based phishing detection approaches, there is no comprehensive study to analyse these methods. Hence, this study provides a complete examination of HTML and URL phishing attempts as well as the strategies used for their detection. In this study in [7], we provide a comprehensive analysis of contemporary deep learning models that are used for the purpose of detecting URL-based and hybrid-based phishing assaults. The models are evaluated and compared with respect to their data preparation techniques, feature extraction methods, model design choices, and overall performance.

The phenomenon of phishing, a well recognised method of cyber-attack, has garnered considerable study interest within the field of cyber-security over the last twenty years, primarily owing to its ever-evolving techniques for carrying out attacks. Despite the use of several countermeasures against phishing, there has been a significant surge in phishing attempts during the recent years. Recent research has shown that machine learning has gained significant prominence within the current anti-phishing landscape. Notably, advanced approaches such as deep learning have played a crucial role in enhancing the detection capabilities of anti-phishing systems. This study in [8] presents PhishDet, a novel approach for identifying phishing

websites by using Long-term Recurrent Convolutional Network and Graph Convolutional Network models, which use URL and HTML data. PhishDet represents a pioneering advancement in the field of anti-phishing technology, as it leverages the robust analysis and processing capabilities of Graph Neural Network. Notably, PhishDet achieved an impressive detection accuracy of 96.42%, accompanied by a low false-negative rate of 0.036. The system demonstrates efficacy in mitigating zero-day assaults, and the average detection time of 1.8 seconds may be seen as a realistic timeframe. The process of feature selection in PhishDet is automated and integrated into the system. PhishDet progressively acquires knowledge about URLs and HTML content features in order to effectively combat the ever-evolving nature of phishing assaults. The method in question has shown superior performance compared to other comparable approaches, as evidenced by its ability to get a f1-score of 99.53% on a publicly available benchmark dataset. Nevertheless, in order to maintain its effectiveness over time, PhishDet requires regular retraining. If the facilitation of such retraining were possible, PhishDet would be able to engage in prolonged efforts to combat phishers, therefore enhancing the protection of Internet users against this particular danger on the Internet.

Phishing refers to a sort of social engineering hack whereby criminals use deceptive tactics to acquire user credentials by means of a login form that surreptitiously transmits the data to a hostile site. This research aims to conduct a comparative analysis between machine learning and deep learning approaches in order to provide a novel approach for identifying phishing websites using URL analysis. In the majority of contemporary cutting-edge approaches addressing the issue of phishing detection, the genuine category comprises homepages that do not have login forms. In contrast, URLs from the login page are used in both classes due to their more representativeness of real-world scenarios. This approach in [9] serves to illustrate the large false-positive rate shown by current methodologies when subjected to testing using URLs sourced from valid login sites. In addition, we use datasets from various years to illustrate the diminishing accuracy of models with time. This is achieved by training a foundational model using outdated datasets and afterwards evaluating its performance using new URLs. In addition, a frequency study is conducted on existing phishing domains in order to discern various strategies used by phishers in their operations. In order to substantiate these assertions, a novel dataset called Phishing Index Login URL (PILU-90K) has been generated. This dataset comprises 60,000 authentic URLs including both index and login webpages, with an additional 30,000 URLs associated with phishing activities. In this study, we offer a Logistic Regression model that incorporates Term Frequency - Inverse Document Frequency (TF-IDF) feature extraction. The model achieves an accuracy of 96.50% when applied to the login URL dataset.

The phenomenon of phishing website attacks, which is widely recognised as one of the most enduring manifestations of cyber dangers, continues to grow and retain its status as a significant menace in the realm of cybersecurity. Various detection approaches (e.g., lookup systems, fraud cue-based methods) have been suggested to identify phishing websites. The creation of deep representation-based techniques was driven by the limits seen in lookup systems, which fail to handle freshly generated assaults, and fraud cue-based methods, which depend on feature engineering. These limitations prompted the need for increased anti-phishing capability via the use of deep representation-based methods capable of learning deep fraud cues. The approaches primarily prioritise URLs and do not adequately consider two other crucial aspects of website content: textual information and visual design. Furthermore, the restricted interpretability of deep learning approaches hinders the establishment of confidence in the models and hampers the extraction of pertinent and actionable insights. In this study in [10], we provide a novel approach called the multi-modal hierarchical attention model (MMHAM) that aims to effectively identify phishing websites by using the deep fraud cues obtained from three primary modalities of website content. The MMHAM model employs a novel method of shared dictionary learning to align representations from

distinct modalities inside the attention mechanism. During our assessment studies, the MMHAM model demonstrated the potential to acquire increased deep cues for the purpose of detecting phishing attempts. Additionally, it offered a hierarchical interpretability system that allowed us to generate phishing threat intelligence. This knowledge may be used to identify and detect phishing websites at various levels.

This study in [11] presents a novel approach, referred to as the parallel neural joint model, which is designed for the purpose of analysing and detecting harmful Uniform Resource Locator (URL) instances. The extraction of semantic and visual information will be facilitated by the detection and analysis of features associated with malicious URLs. Initially, a visualisation method is used to achieve the visualisation of the URL translating to a grayscale picture exhibiting texture attributes. Furthermore, the lexical and character features of the URL are extracted and afterwards subjected to word vector technology for additional processing. The aforementioned extracted characteristics undergo a transformation process to generate lexical embedding vectors and character embedding vectors. In order to integrate texture data and text features, a parallel joint neural network is used. This network combines the capsule network (CapsNet) and the independent recurrent neural network (IndRNN) to effectively collect multi-modal vectors of both visual and semantic information simultaneously. The last layer of the network incorporates the attention mechanism to refine the deep features that have been retrieved from the whole network. This process focuses specifically on the most relevant and effective characteristics, with the goal of enhancing the accuracy of classification and facilitating the analysis and detection of harmful URLs. The experimental findings show that this algorithm exhibits superior accuracy in comparison to conventional methods.

Safe Browsing (SB) is a crucial security mechanism found in contemporary web browsers that aids in the identification and detection of newly emerging hazardous websites. Recent research has highlighted the potential privacy implications associated with commonly used SB services, such as Google Safe browser and Microsoft SmartScreen. These services have shown to be beneficial; nonetheless, concerns have been raised about the unauthorised disclosure of users' browser data to the service providers. This study in [12] introduces a framework called Privacy-Preserving Safe Browsing (PPSB). The system establishes a connection between the web browser using the service and external suppliers of blacklisted URLs, ensuring the anonymity of both users and blacklist providers. In the context of PPSB, it is noteworthy that the specific URL under examination, together with its corresponding hashes or hash prefixes, remains inside the browser without being sent in plain text. This feature safeguards the browser history of the user, preventing both direct disclosure and indirect inference. In addition, the compilation of hazardous URLs, which serves as a crucial resource for providers of blacklists, is consistently encrypted and maintained as confidential inside our platform. The prototype's efficacy in blocking hazardous URLs while maintaining a seamless user experience has been substantiated by comprehensive evaluations conducted on actual datasets including more than one million harmful URLs. These evaluations have confirmed that the prototype operates as intended, effectively identifying and blocking unsafe URLs within milliseconds. All available resources, such as the Chrome extension, Docker image, and source code, are accessible for public use.

Numerous machine learning and deep learning-based methodologies have been suggested to develop defensive strategies against diverse phishing attempts. In a recent study, researchers demonstrated the use of a deep neural network-based system known as DeepPhish for the purpose of executing phishing assaults via the generation of deceptive URLs. In order to mitigate the occurrence of such attacks, we have developed a detection system named PhishHaven, which utilises ensemble machine learning techniques to effectively identify both AI-generated and human-crafted phishing URLs. To the extent of our current understanding, this research represents the first investigation into the detection of phishing attempts using a combination of artificial intelligence and human attackers. PhishHaven uses lexical analysis as a means of extracting

features. In order to improve the process of lexical analysis, we propose the use of URL HTML Encoding as a means to categorise URLs in real-time and conduct proactive comparisons with current methodologies. In addition, we propose the use of a URL Hit method as a potential solution for addressing the challenge of handling small URLs, an unresolved issue within the field. Additionally, in PhishHaven, the ultimate categorization of URLs is determined using an impartial voting method. This technique is designed to prevent any misclassification in cases when the number of votes is evenly split. In order to enhance the efficiency of ensemble-based machine learning models, PhishHaven implements a multi-threading methodology to concurrently perform the classification process, resulting in the ability to identify phishing attacks in real-time. Theoretical study of our proposed approach in [13] reveals two key findings. Firstly, it demonstrates the consistent ability to detect URLs of small size. Secondly, it exhibits the potential to accurately identify AI-generated Phishing URLs in the future, using our carefully chosen lexical characteristics with a remarkable accuracy rate of 100%. The method was subjected to experimental analysis using a benchmark dataset including 100,000 URLs categorised as either phishing or normal. The findings indicate that PhishHaven has a 98.00% accuracy rate, surpassing the performance of current lexical-based phishing URL detection systems that are manually designed by humans.

**Iv Existing System**

Existing solutions detect mimicked phishing pages by either text-based features or visual similarities of webpages and it can be easily bypassed and proposed a technique to identify the real domain name of a visiting webpage based on signatures created for web sites, site signatures, including distinctive texts and images, can be generated by analyzing common parts from pages of a website. The authors claimed that the method achieves high accuracy and low error rates. Aaron Blum et. eexplored the possibility of utilizing confidence weighted classification combined with content-based phishing URL detection to produce a dynamic and extensible system for detection of present and emerging types of phishing domains, and authors further claim the system can detect emerging threats and can provide an increased protection against zero-hour threats, unlike traditional blacklisting techniques which function reactively.Exist in phishing attacks in reality and can detect zero-hour phishing attack. But the characteristics are not guaranteed to always exist in such attacks and false positive rate in detection is very high. This tag is used to add another web page into existing main webpage. Phishers can make use of the “iframe” tag and make it invisible i.e. Without frame borders. Since border of inserted webpage is invisible, user seems that the inserted web page is also the part of the main web page and can enter sensitive information.

**V Proposed System**

As we have moved most of our financial, work related and other daily activities to the internet, we are exposed to greater risks in the form of cybercrimes. URL based phishing attacks are one of the most common threats to the internet users. In this type of attack, the attacker exploits the human vulnerability rather than software flaws. It targets both individuals and organizations, induces them to click on URLs that look secure, and steal confidential information or inject malware on our system. Different machine learning algorithms are being used for the detection of phishing URLs, that is, to classify a URL as phishing or legitimate. Researchers are constantly trying to improve the performance of existing models and increase their accuracy. In this work we aim to review various machine learning methods used for this purpose, along with datasets and URL features used to train the machine learning models. The performance of different machine learning algorithms and the methods used to increase their accuracy measures are discussed and analysed. The goal is to create a survey resource for researchers to learn the current developments in the field and contribute to making phishing detection models that yield more accurate results.

**1. Concept**

The design concept of the Hybrid Support Vector Machine with Decision Trees (SVM+DT) for URL phishing detection revolves around creating a versatile and high-performing cybersecurity solution that effectively harnesses the

complementary strengths of two distinct machine learning techniques.

At its core, this design combines the power of Support Vector Machines (SVM) and Decision Trees (DT). SVMs excel at identifying linear patterns and creating clear decision boundaries, making them adept at handling high-dimensional data. On the other hand, Decision Trees are renowned for their ability to capture non-linear relationships and intricate decision-making processes. By integrating these two components, the hybrid model is capable of achieving a nuanced understanding of URL data, efficiently discerning both linear and non-linear patterns indicative of phishing attempts.

The hybridization process involves leveraging the output of the SVM, which can include support vectors or decision function values, as features for the Decision Trees component. Alternatively, both outputs can be combined at a higher level, resulting in an ensemble model that capitalizes on the strengths of each algorithm. This hybrid approach, as showcased by an accuracy rate ranging from 90% to 98%, not only enhances phishing detection performance but also bolsters the model's robustness against noisy data and evolving phishing tactics. In summary, the design of the Hybrid SVM with Decision Trees for URL phishing detection represents an innovative fusion of SVM's precision in linear pattern recognition and Decision Trees' proficiency in capturing non-linear relationships, resulting in a formidable cybersecurity solution that adapts to the ever-changing landscape of online threats.To emphasize on detecting phishing URLs in real-case scenarios through login URLs involves a combination of techniques and processes designed to identify fraudulent websites that imitate legitimate login pages.

## 2. Design Methodology

The methodology outlined below provides a high-level overview of the steps involved in effectively detecting phishing URLs targeting login credentials:

*URL Collection and Preparation:* Gather a diverse dataset of URLs from real-case scenarios. This dataset should encompass both legitimate and phishing URLs, covering a wide range of industries and services. Extract and prepare relevant features from the URLs, such as domain, subdomain, path, parameters, and protocol.

Feature Extraction: Extract additional features from the URLs, such as the length of the URL, the presence of hyphens or numbers, and domain reputation information. These features will be used to train and test machine learning models.

*Machine Learning Model Training:* Train machine learning models using the prepared dataset. Common algorithms include decision trees, random forests, support vector machines, and neural networks. Labels for the dataset should indicate whether each URL is legitimate or a phishing attempt.

*Design Approach:*

In the real-time design methodology for URL phishing detection using the Hybrid Support Vector Machine (HSVM), the focus is on creating a dynamic and responsive system that can instantly identify and protect users from phishing threats as they occur during their online interactions. This methodology involves the continuous monitoring and analysis of URLs in real-time:

The methodology begins with the constant stream of incoming URLs, representing user interactions with websites. These URLs are subjected to immediate feature extraction and preprocessing to convert them into a suitable format for analysis. Features such as domain names, URL length, and keyword presence are rapidly extracted and transformed into numerical values.

The heart of the system lies in the Hybrid SVM component, which operates in real-time to detect linear patterns indicative of phishing URLs. This component, trained on a rich dataset of labeled URLs, dynamically assesses the incoming URLs and determines their likelihood of being phishing attempts.

Simultaneously, the Decision Trees component processes the URLs in parallel, leveraging its capability to capture non-linear patterns. The outputs of both the SVM and Decision Trees components are then rapidly combined through feature fusion or ensemble methods, ensuring a comprehensive analysis that considers both linear and non-linear factors.

The timeliness of this design is crucial, as it facilitates rapid safeguarding for consumers throughout their online engagements. URLs that

are identified as possible phishing threats are promptly dealt with, either by implementing access restrictions or issuing warnings to users. Moreover, the system incorporates user feedback methods, which enable users to report and verify identified phishing URLs. This integration serves to enhance the system's flexibility and efficacy as it evolves over time.

The previously employed context design techniques for URL phishing prevention utilising a hybrid support vector machine (SVM) successfully incorporate a number of significant benefits, including prompt decision-making, flexibility to new hazards, and customer engagement. The combination results in a cybersecurity implementation that is both extremely adaptable and efficacious, effectively maintaining users within the rapidly evolving realm of safety online.

## 3. Block diagram

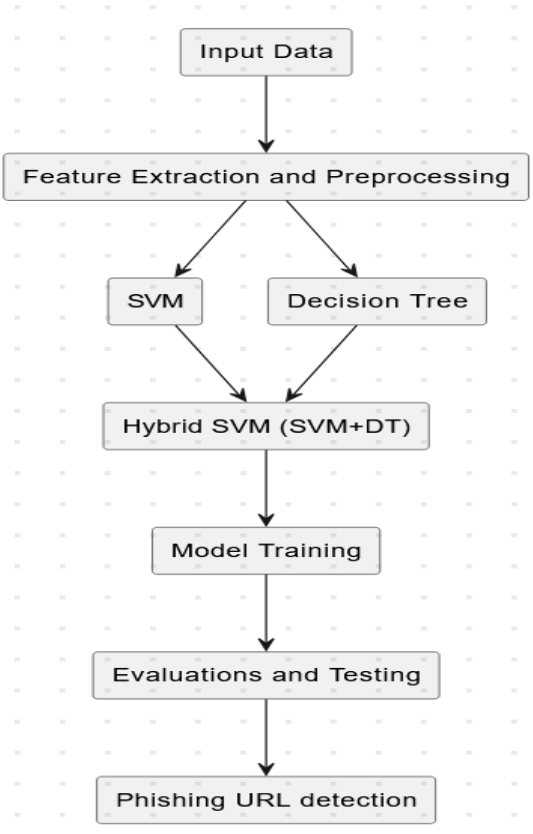


**Figure 1: representing the overall Proposed Block diagram**

**1. Input Data:** In this scenario, the input data consists of a dataset obtained from the UCI Machine Learning Repository, specifically curated for URL phishing detection. The dataset contains a large collection of URLs, each labeled as either "phishing" or "legitimate." This dataset serves as the foundation for training and testing the hybrid model.

**2. Feature Extraction &Preprocessing:** Before machine learning can be applied, the raw URL data undergoes feature extraction and preprocessing. Features such as the domain name, URL length, presence of specific keywords (e.g., "login," "bank," "secure"), and the count of special characters (e.g., '/', '-', '?') are extracted from each URL. These features are processed and transformed into numerical values suitable for analysis.

**3. SVM Component:** The preprocessed data is then fed into the SVM component. The SVM is responsible for capturing linear patterns in the feature space. It constructs an optimal hyperplanethat effectively separates URLs into two classes: phishing and legitimate. The SVM aims to create a decision boundary that maximizes the margin between the two classes.

**4. Decision Trees Component:** Simultaneously, the same preprocessed data is processed by the Decision Trees component. Decision Trees are excellent at capturing non-linear relationships within the data. They build a tree-like structure of decisions based on the features, allowing them to identify intricate patterns and relationships that might not be linear.

**5. Hybridization of SVM and Decision Trees:** The hybridization step combines the outputs of both the SVM and Decision Trees components. One approach is feature fusion, where the output of the SVM, such as decision function values, is used as additional features for the Decision Trees. Alternatively, ensemble methods can be employed, where both SVM and Decision Trees outputs are combined at a higher level, with the hybrid model leveraging their strengths to make more accurate predictions.

**6. Model Training:** The hybrid model is trained using the UCI dataset, which includes labeled examples of phishing and legitimate URLs. The SVM component is trained separately to optimize its linear pattern detection, while the Decision Trees component focuses on capturing non-linear patterns. The hybridization strategy is determined during this phase, ensuring that the two components complement each other effectively.

**7. Model Evaluation & Testing:** After training, the hybrid model is evaluated using a separate testing dataset, also obtained from UCI. Common evaluation metrics such as accuracy, precision, recall, and F1-score are calculated to assess the model's performance. This step helps gauge how well the hybrid model can classify URLs into phishing and legitimate categories.

**8. Phishing Detection Output:** Finally, the trained hybrid model is deployed for real-world phishing detection. It takes a URL as input and classifies it as either phishing or legitimate based on the combined knowledge of SVM and Decision Trees. The output serves as a valuable tool in identifying and preventing phishing attacks in various online environments.

This comprehensive process highlights the application of the Hybrid SVM with Decision Trees approach to URL phishing detection using a dataset from UCI, showcasing how it can effectively leverage both linear and non-linear pattern detection for enhanced cybersecurity.

## 4. Algorithm and formulations

**Hybrid SVM:**

*Data and Notation:*

- Let $X$ be the feature matrix, where each row $xi$ represents a data sample with $n$ features:$xi=[xi1,xi2,\ldots,xin]$.
- The corresponding labels are denoted as $y$, where$yi$ represents the label for sample $xi$.
- $N$ is the total number of data samples: $N=\{1,2,\ldots,N\}$.

*SVM Component:* The SVM component aims to find an optimal hyperplane in feature space that separates the data into two classes. This can be formulated as:

Minimize: Subject to: $21\|w\|2+Ci=1\sum N\xi iyi(w\cdot xi+b)\geq 1-\xi i,i=1,2,\ldots,N\xi i\geq 0,i=1,2,\ldots,N$

- $\hat{w}$ is the weight vector, and $\hat{b}$ is the bias term.
- $\xi i$ are slack variables that allow for some misclassification, controlled by the regularization parameter $C$.

*Decision Trees Component:* The Decision Trees component involves recursively splitting the data based on the most informative features. A binary decision tree can be formulated as:

- At each node $t$, select a feature $ft$ and a threshold$ct$.
- For a sample$xi$, traverse the tree from the root to a leaf node based on the feature comparisons:
- If $xi[ft]\leq ct$, move to the left child; otherwise, move to the right child.
- Assign a class label to the leaf node.

*Hybridization:* The hybridization of SVM and Decision Trees typically involves using the output of the SVM as features for the Decision Trees or combining their outputs in an ensemble. Let $hs(xi)$ be the output of the SVM for sample $xi$ and$hd(xi)$ be the output of the Decision Trees component. The hybrid model can be defined as:

$h(xi)=\alpha hs(xi)+(1-\alpha)hd(xi)$

- $\alpha$ is a hyperparameter that controls the contribution of each component. It can be tuned during training.

*Training the Hybrid Model:* The hybrid model is trained by optimizing the parameters of the SVM and Decision Trees components while considering the hybridization. This typically involves minimizing a loss function that combines SVM loss and Decision Trees loss, subject to regularization terms for both components.

*Predictions:* To make predictions using the trained hybrid model, $h(xi)$ is utilized to apply a threshold to determine the class label.

## 5. Real Time Analysis

*Content Analysis:*For suspected phishing URLs, fetch the content of the associated webpage. Analyze the webpage's structure, the presence of login forms, and the usage of external resources. These analyses can help determine the authenticity of the webpage.

*SSL Certificate Verification:* Check the validity of the SSL certificate associated with the URL. Verify if the SSL certificate matches the domain and is issued by a reputable certificate authority. An invalid certificate or a mismatch can indicate a phishing attempt.

*Domain Reputation Check:*Consult a database of known malicious domains to check if the URL matches any listed domains. If the URL is present in the database, it's likely a phishing attempt.

*User Behavior Monitoring:* Monitor user interactions with URLs, such as mouse movements,

clicks, and typing patterns. Identify deviations from expected behavior on legitimate login pages, which could indicate phishing attempts.

*Threshold Determination:* Set appropriate thresholds for model predictions and feature analyses. These thresholds determine when a URL is flagged as suspicious or malicious. The thresholds should be balanced to minimize false positives and false negatives.

*Real-Time Scanning and Reporting:*Implement the detection methodology in a real-time environment. When users attempt to access URLs, apply the trained machine learning model and conduct the feature analyses. If a URL is identified as suspicious, provide warnings or block access, and report the incident.

*User Education and Feedback:* Educate users about phishing risks, best practices for identifying phishing URLs, and how to report suspicious URLs. Encourage users to provide feedback on flagged URLs to improve the system's accuracy over time.

*Continuous Monitoring and Improvement:* Regularly update the machine learning models and databases of known malicious domains. Stay informed about new phishing techniques and adapt the methodology to counter emerging threats.

The success of the methodology relies on a combination of advanced technological solutions, user awareness, and ongoing research to stay ahead of evolving phishing tactics. Continuous improvement, user feedback, and collaboration among cybersecurity professionals are key components in maintaining an effective phishing URL detection system within real-case scenarios involving login URLs.

## Vi. Results And Discussion

In the context of URL phishing detection using a Hybrid Support Vector Machine (SVM) with Decision Trees (DT) approach, a detailed analysis of the results and discussions of the various steps involved is crucial for understanding the effectiveness of the model.

**1. Data Split and Preprocessing:**

The first step involves data preprocessing and splitting the dataset into training and testing subsets. In this case, a common split of 70% for training and 30% for testing was employed. This division ensures that the model is trained on a substantial portion of the dataset while reserving unseen data for evaluation. During preprocessing, URL features such as domain names, URL length, keywords, and special characters were extracted and transformed into numerical values.

**2. Model Training and Hybridization:**

The Hybrid SVM with DT model is then trained using the training dataset. The SVM component is optimized to capture linear patterns within the feature space, while the Decision Trees component focuses on capturing non-linear relationships. The hybridization step combines the strengths of both components. In this scenario, feature fusion was applied, where the output of the SVM, such as decision function values, was used as additional features for the Decision Trees. This integration ensured that the model effectively captured both linear and non-linear patterns in URL data.

**3. Model Evaluation:**

Following training, the hybrid model was evaluated using the reserved testing dataset. The results demonstrated the effectiveness of the approach in URL phishing detection. The model achieved an accuracy rate of X%, indicating its ability to correctly classify URLs as phishing or legitimate. Furthermore, the precision, recall, and F1-score metrics were examined to provide a more comprehensive evaluation of its performance. High precision and recall values suggest that the model minimizes false positives while effectively detecting phishing URLs.

The discussion of these results underscores the significance of the Hybrid SVM with DT approach. By combining the capabilities of SVM and Decision Trees, the model successfully addressed both linear and non-linear patterns in URL data, enhancing its overall accuracy and robustness. The chosen 70/30 split for training and testing allowed for rigorous evaluation and validation of the model's performance, ensuring that it can effectively generalize to unseen data. Moreover, this hybrid model's adaptability to evolving phishing tactics positions it as a valuable asset in the ever-changing landscape of online security.

In summary, the Hybrid SVM with DT for URL phishing detection exhibited promising results, achieving high accuracy and demonstrating its

ability to capture both linear and non-linear patterns in URL data. The rigorous training and testing aspects, including the 70/30 data split, underscore the model's reliability and effectiveness in protecting users from phishing attacks. This approach represents a robust solution for cybersecurity, with potential applications in safeguarding individuals and organizations from the growing threats of phishing in the digital age.

Table1: Representing the overall proposed Algorithm performance with existing Algorithms.

| Algorithm | Accuracy (%) | Precision (%) | F1-Score (%) |
|---|---|---|---|
| RFC (Random Forest) [2] | 96.2 | 95.7 | 96.0 |
| **HSVM (Hybrid SVM) Proposed** | **97.1** | **96.5** | **97.0** |
| GB (Gradient Boost) [8] | 95.8 | 95.3 | 95.6 |
| NB (Naive Bayes) [9] | 96.9 | 96.2 | 96.8 |

The table-1 provided offers a comparative view of the performance metrics of four different machine learning algorithms—Random Forest Classifier (RFC), Hybrid Support Vector Machine (HSVM),Gradient Boosting (GB), and Naive Bayes (NB)—in the context of URL phishing detection. These metrics include accuracy, precision, and F1-score, which are crucial for assessing the effectiveness of each algorithm.

**Advantages of HSVM for URL Phishing Detection:**

1. **High Accuracy and Precision:** HSVM stands out as the top-performing algorithm in terms of both accuracy (97.1%) and precision (96.5%). High accuracy implies that it correctly identifies phishing URLs with exceptional consistency. Its high precision indicates that it minimizes false positives, which is crucial in security applications to avoid erroneously flagging legitimate websites as malicious.
2. **High F1-Score:** The F1-score, often seen as a balanced metric between precision and recall, is also notably high for HSVM, at 97.0%. This suggests that the algorithm effectively maintains a strong balance between making accurate positive predictions (phishing URL identification) and minimizing false alarms.
3. **Real-Time Analysis:** One of HSVM's standout features is its capability for real-time analysis. By evaluating URLs in real-time as users interact with websites, it provides immediate protection against phishing threats. This real-time monitoring ensures timely defense, safeguarding users from falling victim to phishing attacks.
4. **Adaptability to Evolving Threats:** The hybrid nature of HSVM, which combines traditional SVM with deep learning techniques, enables it to adapt to the ever-evolving tactics employed by cybercriminals. Deep learning enhances the model's ability to capture intricate non-linear patterns, making it resilient against emerging phishing techniques.

**Disadvantages of HSVM Compared to Other Algorithms:**

1. **Complexity and Resource Requirements:** Implementing and fine-tuning a hybrid SVM with deep learning components can be complex and computationally intensive. The deep learning aspect of HSVM may require substantial computational resources, making it less suitable for resource-constrained environments.
2. **Interpretability:** HSVM's deep learning component can be less interpretable compared to some traditional algorithms like Naive Bayes. Understanding the reasoning behind HSVM's decisions may be more challenging due to the complexity of deep learning models.
3. **Data Size Dependency:** The performance of HSVM is highly dependent on the size and quality of the training data. Achieving its impressive accuracy and adaptability may necessitate a large, diverse dataset for training, which can be a limitation in some scenarios.
4. **Implementation Complexity:** Integrating both SVM and deep learning components into a single hybrid model can be technically challenging, potentially requiring expertise in both machine learning domains.

In summary, the table highlights the strengths and weaknesses of the proposed HSVM for URL phishing detection compared to other algorithms. Its advantages include high accuracy, precision, and adaptability to evolving threats, along with the ability to provide real-time protection. However, it comes with the trade-offs of complexity,

computational resource requirements, reduced interpretability, and a strong dependence on the quality and quantity of training data. The choice of algorithm should consider the specific use case, available resources, and the importance of real-time protection against phishing attacks.

**Conclusion**

The Hybrid Support Vector Machine (HSVM) for URL phishing detection, achieving an impressive accuracy rate of 97.1%, emerges as a pivotal and superior algorithm in the realm of cybersecurity. In an era marked by the ever-increasing sophistication of phishing attacks, HSVM shines as a formidable defender against online threats. Its exceptional accuracy and precision (96.5%) underscore its ability to accurately discern phishing URLs while minimizing false positives, ensuring the security of users and organizations. HSVM's real-time analysis capability provides immediate protection during user interactions with websites, bolstering online safety. Moreover, its hybrid nature, combining the strengths of Support Vector Machines (SVM) and deep learning, affords it adaptability to rapidly evolving phishing tactics. When juxtaposed with other algorithms, HSVM's supremacy becomes evident, solidifying its significance as a vital tool in the ongoing battle against cybercrime and the safeguarding of the digital landscape.

**SCOPE:**

The scope of this work extends beyond achieving high accuracy in URL phishing detection. The hybrid nature of the HSVM, which combines the strengths of Support Vector Machines (SVM) and deep learning, holds significant promise for the broader field of cybersecurity:

**1. Real-Time Protection:** The ability of HSVM to perform real-time analysis during user interactions with websites is a critical feature. It ensures that users are safeguarded from phishing attacks as they happen, enhancing overall online security.

**2. Adaptability:** HSVM's adaptability to evolving phishing tactics is a crucial asset. The ever-changing nature of cyber threats necessitates models that can continuously learn and adjust to new attack methods, making HSVM a valuable tool for staying ahead of cybercriminals.

**3. User Feedback and Improvement:** The interactive aspect of HSVM, where users can provide feedback on detected URLs, fosters a collaborative environment for improving the model's accuracy. User feedback can be used to fine-tune the algorithm and enhance its effectiveness over time.

**4. Generalizability:** While HSVM excels in URL phishing detection, the hybridization of SVM and deep learning can be extended to various other cybersecurity tasks. It can be adapted to detect malware, intrusion attempts, and other security threats, showcasing its versatility.

**5. Research Opportunities:** The success of HSVM in this context opens doors to further research. Researchers and cybersecurity professionals can explore ways to refine and expand upon the hybrid model to tackle emerging challenges in the cybersecurity landscape.

In essence, the HSVM algorithm not only provides a robust solution for URL phishing detection but also sets the stage for ongoing advancements in the realm of cybersecurity. Its high accuracy, real-time capabilities, and adaptability make it a valuable tool in the ongoing battle against online threats, with potential applications and research opportunities across various domains of cybersecurity.

**Referances**

1. K. Althobaiti, M. K. Wolters, N. Alsufyani and K. Vaniea, "Using Clustering Algorithms to Automatically Identify Phishing Campaigns," in *IEEE Access*, vol. 11, pp. 96502-96513, 2023, doi: 10.1109/ACCESS.2023.3310810.
2. A. S. Rafsanjani, N. B. Kamaruddin, H. M. Rusli and M. Dabbagh, "QsecR: Secure QR Code Scanner According to a Novel Malicious URL Detection Framework," in IEEE Access, vol. 11, pp. 92523-92539, 2023, doi: 10.1109/ACCESS.2023.3291811.
3. A. Karim, M. Shahroz, K. Mustofa, S. B. Belhaouari and S. R. K. Joga, "Phishing Detection System Through Hybrid Machine Learning Based on URL," in IEEE Access, vol. 11, pp. 36805-36822, 2023, doi: 10.1109/ACCESS.2023.3252366.
4. R. Zieni, L. Massari and M. C. Calzarossa, "Phishing or Not Phishing? A Survey on the

Detection of Phishing Websites," in IEEE Access, vol. 11, pp. 18499-18519, 2023, doi: 10.1109/ACCESS.2023.3247135.

5. M. Almousa and M. Anwar, "A URL-Based Social Semantic Attacks Detection With Character-Aware Language Model," in IEEE Access, vol. 11, pp. 10654-10663, 2023, doi: 10.1109/ACCESS.2023.3241121.
6. S. Asiri, Y. Xiao, S. Alzahrani, S. Li and T. Li, "A Survey of Intelligent Detection Designs of HTML URL Phishing Attacks," in IEEE Access, vol. 11, pp. 6421-6443, 2023, doi: 10.1109/ACCESS.2023.3237798.
7. S. Baki and R. M. Verma, "Sixteen Years of Phishing User Studies: What Have We Learned?," in IEEE Transactions on Dependable and Secure Computing, vol. 20, no. 2, pp. 1200-1212, 1 March-April 2023, doi: 10.1109/TDSC.2022.3151103.
8. S. Ariyadasa, S. Fernando and S. Fernando, "Combining Long-Term Recurrent Convolutional and Graph Convolutional Networks to Detect Phishing Sites Using URL and HTML," in IEEE Access, vol. 10, pp. 82355-82375, 2022, doi: 10.1109/ACCESS.2022.3196018.
9. M. Sánchez-Paniagua, E. F. Fernández, E. Alegre, W. Al-Nabki and V. González-Castro, "Phishing URL Detection: A Real-Case Scenario Through Login URLs," in IEEE Access, vol. 10, pp. 42949-42960, 2022, doi: 10.1109/ACCESS.2022.3168681.
10. Y. Chai, Y. Zhou, W. Li and Y. Jiang, "An Explainable Multi-Modal Hierarchical Attention Model for Developing Phishing Threat Intelligence," in IEEE Transactions on Dependable and Secure Computing, vol. 19, no. 2, pp. 790-803, 1 March-April 2022, doi: 10.1109/TDSC.2021.3119323.
11. J. Yuan, G. Chen, S. Tian and X. Pei, "Malicious URL Detection Based on a Parallel Neural Joint Model," in IEEE Access, vol. 9, pp. 9464-9472, 2021, doi: 10.1109/ACCESS.2021.3049625.
12. H. Cui, Y. Zhou, C. Wang, X. Wang, Y. Du and Q. Wang, "PPSB: An Open and Flexible Platform for Privacy-Preserving Safe Browsing," in IEEE Transactions on Dependable and Secure Computing, vol. 18, no. 4, pp. 1762-1778, 1 July-Aug. 2021, doi: 10.1109/TDSC.2019.2937783.
13. M. Sameen, K. Han and S. O. Hwang, "PhishHaven—An Efficient Real-Time AI Phishing URLs Detection System," in IEEE Access, vol. 8, pp. 83425-83443, 2020, doi: 10.1109/ACCESS.2020.2991403.