PHISHING DETECTION SYSTEM THROUGH HYBRID MACHINE LEARNING BASED ON URL.

Ms. Sangeetha Raj S *, Vikas Reyaz Bhat **, Sumiksha Bhadwal ***, Shiva Singh Shishodia ****, Puneet Reddy *****

*(Faculty of Computer Science and Engineering, AMC Engineering College, Bangalore Email: sangeetha.rajasekar@amceducation.in)

**(UG Student of Computer Science and Engineering, AMC Engineering College, Bangalore Email : vikas.22cs422@ amceducation.in)

***(UG Student of Computer Science and Engineering, AMC Engineering College, Bangalore Email: sumikshabhadwal999@gmail.com)

****(UG Student of Computer Science and Engineering, AMC Engineering College, Bangalore Email: shivasinghshishodia@gmail.com)

*****(UG Student of Computer Science and Engineering, AMC Engineering College, Bangalore Email: puneetreddy0912@gmail.com)

This work was supported by Student's Final year Project 2025 From AMC Engineering College Bangalore.

ABSTRACT: Currently, numerous types of cybercrime are organized through the internet. Hence, this study mainly focuses on phishing attacks. Although phishing was first used in 1996, it has become the most severe and dangerous cybercrime on the internet. Phishing utilizes email distortion as its underlying mechanism for tricky correspondences, followed by mock sites, to obtain the required data from people in question. Different studies have presented their work on the precaution, identification, and knowledge of phishing attacks; however, there is currently no complete and proper solution for frustrating them. Therefore, machine learning plays a vital role in defending against cybercrimes involving phishing attacks. The proposed study is based on the phishing URL-based dataset extracted from the famous dataset repository, which consists of phishing and legitimate URL attributes collected from 11000+ website datasets in vector form. After pre-processing, many machine learning algorithms have been applied and designed to prevent phishing URLs and provide protection to the user. This study uses machine learning models such as decision tree (DT),linear regression (LR), random forest (RF), naive Bayes (NB), gradient boosting classifier (GBM), K-neighbours classifier (KNN), support vector classifier (SVC), and proposed hybrid LSD model, which is a combination of logistic regression, support vector machine, and decision tree (LR+SVC+DT) with soft and hard voting, to defend against phishing attacks with high accuracy and efficiency.

INDEX TERMS: Voting classifier, ensemble classifier, machine learning, uniform resource locator (URL), logistic regression, support vector machine, and decision tree (LSD), protocol, cyber security, social networks**.**

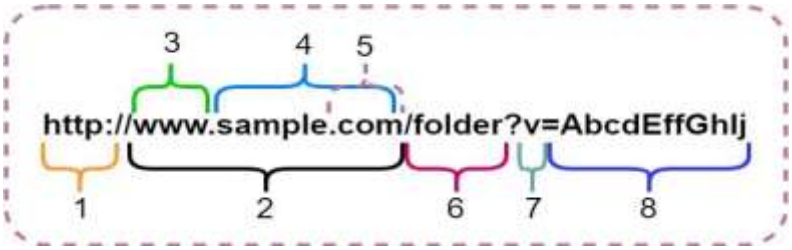

FIGURE 1. URL presentation based on HTTP.

I. INTRODUCTION: Phishing is considered one of the finest prevalent cyber threats today, targeting individuals and organizations to steal data such as login credentials, financial details, or personal data. Such attacks are generally conducted through deceptive websites, emails, or messages designed to mimic legitimate entities. Despite advancements in cyber security, phishing continues to evolve, employing sophisticated techniques that make detection increasingly challenging. This highlights the acute need for advanced detection systems capable of identifying phishing attempts in real time. This project addresses the phishing problem by proposing a multi algorithmic detection platform that integrates artificial intelligence techniques with advanced text processing methods. Specifically, the system uses TFIDF vectorization to extract and represent features from textual data such as URLs, email content, or web page descriptions. TFIDF is a widely used method in Human Language Technology that assigns importance to words based on their frequency and uniqueness within a dataset, making it ideal for identifying patterns in phishing attempts. To classify the extracted features, three next generation AI algorithms employed: SVM, Gradient Boosting, and RF. SVM is known due its ability to binary classification tasks, making it suitable for distinguishing phishing from legitimate data. Gradient Boosting is included for its ability to showcase complex ties in data, while Random Forest is chosen for its robustness and ensemble-based approach to improve classification accuracy. The project further explores combining the outputs of these algorithms to enhance performance through ensemble techniques. By leveraging the strengths of multiple models, the system aims to reduce errors, improve detection accuracy, and address diverse phishing patterns. The suggested method is tested on phishing datasets to evaluate its scalability, efficiency, and real-world applicability. With a focus on creating a user friendly and reliable framework, this project aims to contribute significantly to cyber security by providing a practical tool for combating phishing attacks. To address these challenges, machine learning (ML) has been widely adopted for phishing detection. ML models can learn from patterns and characteristics of phishing URLs to accurately classify them as legitimate or malicious. However, relying on a single algorithm often leads to limitations in accuracy and increases the likelihood of false positives. To address these challenges, machine learning (ML) has been widely adopted for phishing detection. ML models can learn from patterns and characteristics of phishing URLs to accurately classify them as legitimate or malicious. Phishing detection has long been a focus of cybersecurity research and product development, with numerous systems proposed and deployed to combat fraudulent websites. Traditional systems typically employ static defense mechanisms such as blacklisting and heuristic rule based methods. Blacklists work by maintaining a repository of previously identified phishing URLs and domains, denying access when a match is found. However, due to the dynamic and fast-evolving nature of phishing attacks, blacklists quickly become outdated and fail to detect zero-day threats—newly launched phishing sites that haven't yet been reported. As phishing strategies change, the system can adapt new features and retrain models to make sure that it is still effective even as attackers.

The growing sophistication of phishing attacks has rendered traditional detection systems ineffective, and hence there is a need for sophisticated solutions. To determine the disadvantages of phishing detection systems currently in use, this project introduces a Multi Algorithmic Approach that enhances phishing detection accuracy, scalability, and adaptability, rendering it able to respond to changing phishing threats. In contrast to traditional detection systems that tend toon static lists of known phishing URLs or on a single detection algorithm, this proposedsystem employs AI methods and sophisticated text processing methods to provide better performance.A major part of this proposed solution is the utilization of TFIDF vectorization (Term Frequency Inverse Document Frequency), a robust feature extraction method. TFIDF examines the significance of words in a dataset, determining useful terms that assist separate phishing from genuine content. This technique is especially valuable in phishing detection because phishing attacks typically include subtle manipulation of textual content, such as deceptive domain names, fishy wording, or misleading language. By converting the text data into the representation that represent the significance of each word, TFIDF allows the detection system to identify patterns suggestive of phishing, even when known keywords are not used. The boosting step enables to glance on hard to categorize examples, enhancing the overall accuracy.

Random Forest: Random Forest is also utilized to generate decision trees and gives the classification or mean prediction (regression). It is famous for its stability and capability to work with noisy data, which is very important in phishing detection since phishing attacks are mostly based on intentional obfuscations and deceptive patterns. Through the fusion of AI methodologies, the system exploits the strength of each one, enhancing robustness and generalizability in the detection process. By averaging different techniques, the system find the threat may be bypassed by an individual algorithm,resulting in increased detection and reduced false alarms. The use of multi-algorithmic mechanisms also allows the system to be adaptable in response to new phishing tactics because the models can be retrained using new information to remain current with new phishing trends. Additionally, scalability of this suggested system is an important strength over current detection systems. Phishing detection models traditionally tend to struggle to scale well as the volume of phishing attacks grows or as new methods are developed. The use of several machine learning models and feature extraction techniques enables the system to process big data for phishing attacks are continuously on the increase. Moreover, the adaptability of the system ensures that it can seamlessly integrate new phishing trends without needing a total overhaul, hence offering a sustainable solution to continuous cybersecurity issues. Unlike conventional list-based systems, which rely on pre-established lists of phishing URLs or domains, this method provides a dynamic and proactive solution.

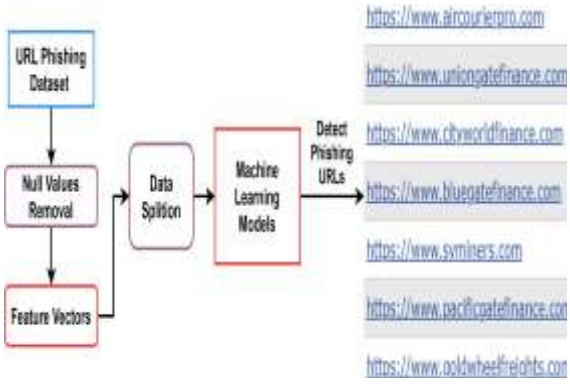

FIGURE 2. Detection of phishing URLs and structure of proposed approach.

II. RELATED WORK:

Phishing is the most important problem in network Internet,domains.Numerous,research-ers have tried to offer facilities to save users from cyber-attacks by stopping the phishing of URLs using machine learning, deep learning, black lists, and white lists. Two types of phishing detection systems have been proposed and used in earlier research: list-based and machine learning identification systems. This is a two-part section:earlier list-based and machine-learning-based researches.

A. LIST BASED IDENTIFICATION OF SYSTEM:

List-based phishing identification systems utilize two different lists white lists and blacklists for the mapping and categorization of approved and phishing webpages.Whitelist based on the as per identification systems generate secured and safe websites to generate the required information. A suspicious website only has to resemble the website of the whitelists; if it is not on the whitelist, it is considered suspicious and threatened as by the user. In [20]. To create a whitelist-based system that produces a whitelist by observing and tracking the IP address of all websites that have the login interface for the end-user utilized by the users to input their details. When the user makes use of this login interface, the Windows 2008 system shows an alert for registered information details incompatibility. This is why this system mechanism suspects genuine sites browsed by users for the first time. Reference [21] proposed a system that warns users regarding a phish ing website by maintaining and updating the whitelist periodically and automatically. The performance of this system is dependent on two factors: attribute extraction buried in the relation between the source code and the module that is compatible with the IP address of the domain. Based on the initial conclusions, 86.02 the true positive rate was 1.48% false negative score was this research.Blacklist Were Gathered as per the according to the history of URLs termed as phishing websites. Many sources, including user notifications, spam detection systems, and third-party authorities, are employed to gather record entries for list making. Blacklist allows systems to block attack ers from capturing their IP address and URLs. Thus, n ext time attackers have to utilize a different URL or IP address since blacklist-based system picks up their old URLs or IPs. A System security administration can automatically refresh the blacklist every so often to fend off new attackers by detecting bad URLs or IPs. Alternatively, individuals can download the lists for updating the security system. Zero day attacks primarily impact systems since blacklist-based systems cannot identify a new or day-one attack. These intrusion detection systems have a lower false-positive score compared to machine learning-based systems. The detection accuracy of intrusions or attacks of these systems based on the blacklist is highly accurate, and with success rate of around 20%, as stated by [22] and [23]. Consequently, this indicates that some companies' identification systems founded on blacklist mechanisms, like Phish.Net [24] and Google Safe Browsing API [25], are trustworthy for detection of phishing attacks based on blacklists. Approximate matching algorithms are employed by such security systems in order to match malicious URLs with URLs in the blacklist. Blacklists using these systems must be updated frequently. Besides, the rapid growth in blacklists requires lavish system support [26], [27].

B. MACHINE LEARNING BASED IDENTIFICATION SYSTEM.

Machine learning is the most popular technique for identifying malicious and suspicious websites by using URLs. Classification of phishing URLs is an important domain in machine learning.

in machine learning. A lot of data features are needed to obtain machine-learning-based security systems and to train the model on features that are correlated with legitimate and phishing website labels. The excellent performance of machine learning algorithms enables them to detect easily hidden or first-time attacks that are not on a blacklist. The authors [28] created a phishing detection system based on text classification called CANTINA.This method retrieves features as keywords with the help of a feature extraction method referred to as term frequency inverse document frequency(TFIDF). These retrieved keywords were employed to query the Google search engine, and if any of these websites were discovered, they were labeled as valid websites. Nevertheless, the success of this research is limited

since it is highly sensitive to English vocabulary. Then, another improved methodology was suggested by [29], relying on the properties of 15 various HTMLs, CANTINA+. The best precision of 92% was yielded by this model, which resulted in an astonishing number of false-positive predictions. Reference [30] designed an anti-phishing-based protection system named Phish WHO, which entails three levels in order to tell whether a website is original. The first level includes a procedure in order to gather keywords to discern malicious websites, and second-level keywords are employed to determine potential asso

ciated domains through a search engine.The victim domain was separated by using the features acquired from these websites. Lastly, at the bottom level, the system determines if the website with doubts at the bottom level is authorized. In 2011, [31] suggested a system for the identification of phishing websites by categorizing these websites by utilizing the number of attributes, such as directory, file name, domain name, counting the number of special characters, and length. By applying a support vector machine (SVM), security systems can classify phishing websites in offline mode. Other techniques and machine learning algorithms, such as

weighted confidence, adaptive regularization of weights, and online perceptrons, are adopted for classification in online mode. In the analyses of the comparative results, the experiments indicate that the adaptive weights regularization algorithm performs better than the other algorithms by having the

highest rate of accuracy and using the minimum amount of system resources. The message title and ranking based on

the incoming message ranked as stated in Islam and Abawajys study [32]. These researches generated a classification system based on the use of more than one layer to explain the importance of messages. The experimental results revealed that the proposed method decreased the number of false positives. The discriminant features are extracted by [33] and associated with the security of the transport layer synchronically among

features of attributes that are based on URLs, such as the total number of used slashes, length, positions of dots, and

numbers in the subdomain and URL names. The Apriori algorithm was set based on rule discovery utilizing rule mining.The experimental results showed that 93%accurately identified the phishing URLs.

Literature	Summary	Pros.	Cons.
[41].	Email based Phishing Detects system using machine learning and NLP techniques.	The major advantage is that the NLP is used to detect the appropriate sentences.	It depends on The email text content analyses. ML is utilizing in the creation of blacklist based on pairs of malicious keywords.limited dataset of 5,009 from phishing and 5,000 from legitimate emails.
[42]	Proposed an entropy based collaborative mechanism for early detection of low rate and high rate DDOS attack and flash events. Packet Header, Time Window size, and other generalized parameters	CAIDA, MIT Lincoln, and FIFA	F measure, precision, False Positive rate and accuracy
[29].	The rich machine learning based system is implemented to detect the phishing websites and URLs based on contents	The main is to catch the novel phishing URLs based on frequently evolving attacks. They expands the number of features for URLs attributes from their previous work (Zhang,2007).	4883 legitimate and 8110 phishing website based limited dataset was used. use services of the third-party companies. use 100 site data collected belongs to only English language and location-specific.
[40]	The machine learning based detection of phishing attack on the client-side through web pages. The Principal Component Analyses (PCA) used with the Random forest classifier to classify the combined image analyses and heuristic feature based analyses.	It is not dependent on the services of third parties and provide detection in real time. The high accuracy achieved in detection. independence from language. achieve highest accuracy in detection. also check the web page is replaced with the image or not and detect phishing.	this system needs to first download the complete page. also used services of third-party. using limited 11,055 data, 55.69% belong to phishing, 30 features used which are address bar, abnormal, HTML, javascript and domain based features. The dataset is limited and consists of 3,717 malicious and 3,640 legitimate URLs.
[39]	The combined approach is proposed by utilizing the neural network and reinforcement learning techniques to detect the phishing in emails.	It fast in detecting phishing emails before the end user saw it. does not dependent on services of third party. provide detection of real time.	The services of the third party are used like the domain age. The dataset is limited with the number of 1,400 data and 17 number of features. The limited dataset is used with 11,055 legitimate and phishing websites and dependent on third-party services with 20 features
[34]	The non linear regression on the bases of a meta-heuristic algorithm by using two methods of feature selection such as wrapper and decision tree.	features based on the NLP. The 3 different machine learning algorithms are used and also used hybrid features. 7% increased performance in comparison of Buber, 2017a. 278 features which are consists of 40 NLP and 238 word features. This system was implemented based on adaptive techniques in producing the network. Provide the language based independence.	this system needs to first download the complete page. also used services of third-party. using limited 11,055 data, 55.69% belong to phishing, 30 features used which are address bar, abnormal, HTML, javascript and domain based features. The dataset is limited and consists of 3,717 malicious and 3,640 legitimate URLs.
[33]	Define some URL features, and with them, they generate some rules with apriori and predictive apriori rule generation algorithms.	The original repository of datset UCI is decreased from 30 to 20 number of features that will helps in achieving the better outcome with the methods of decision trees. fast detection with rules (especially with apriori rules)	
[36]	Uses NLP for creating some features and with the use of these features classifies the URLs by using three different machine learning approach.	It fast in detecting phishing emails before the end user saw it. does not dependent on services of third party. provide detection of real time.	It is not dependent on the services of third parties and provide detection in real time. The high accuracy achieved in detection. independence from language. achieve highest accuracy in detection. also check the web page is replaced with the image or not and detect phishing.
[34]	The non linear regression on the bases of a meta-heuristic algorithm by using two methods of feature selection such as wrapper and decision tree.	These systems are appropriate for the client side employment. THese are online classification based system. Resilient to noisy data training.	The services of the third party are used like the domain age. The dataset is limited with the number of 1,400 data and 17 number of features. The limited dataset is used with 11,055 legitimate and phishing websites and dependent on third-party services with 20 features
[33]	Define some URL features, and with them, they generate some rules with apriori and predictive apriori rule generation algorithms.	Its not dependent on the services of third parties. provide real-time detection. Enhance the rate of accuracy and the detection stability. able to detect novel phishing websites also known as zero-day attack.	

III. SYSTEM FLOW EXPLAINATION:

The Phishing URL Detection System uses a systematic flow that begins with input data and ends with the notification of users about possible threats. The system uses sophisticated methods such as TF-IDF vectorization, machine learning algorithms, and real-time monitoring to make effective phishing detection and response possible. The system for detecting phishing uses a systematic flow that involves several components in order to guarantee correct detection, real-time monitoring, and timely user warnings. The system starts by taking user input in the form of a URL and then processes the input through a series of machine learning and feature extraction steps. Through TF-IDF vectorization, the system transforms the text-based URL into a numerical format for analysis. This processed data is then passed through various classifiers—SVM, Random Forest, and Gradient Boosting—to assess if the URL is phishing or not. After individual predictions are generated, a model aggregation method aggregates their results to create a final decision. This output is displayed in real-time graphs for trend and performance monitoring. If a phishing URL is identified, the system immediately sends an SMS notification to alert the user. Every step of the flow is optimized to improve detection accuracy, system performance, and user safety, offering a strong and proactive safeguard against phishing attacks. After the classification, the system initiates real-time visualization, where graphical displays of detection processes are shown on a web interface. This enables administrators or users to track phishing patterns over time and be informed about potential threats. Concurrently, if a phishing URL is detected, the system immediately sends the SMS notification mechanism, sending messages to the affected user or admin. The quick communication aids in fast preventive measures, hence less risk of damage from phishing attacks. Overall, the system flow is made modular, scalable, and responsive—protecting users in real-time and keeping the detection process effective even against changing phishing patterns. Every module in the flow is interdependent and is part of building a unified, smart defense against phishing attacks. Users or automated systems at the start of the process enter URLs into the detection system. URLs may be from various sources such as websites, or automated crawlers. Both individual user submissions and batch inputs from other systems are handled by the system so that there is flexibility in operation. Data Source: URLs may be entered directly input by users, gathered from web traffic, or derived from email links. Purpose: This step prepares the URLs for analysis in the later phases of the detection process. This proactive notification system significantly enhances user safety and system responsiveness, reducing the effect of phishing attacks by allowing real-time response and defense. All of these elements—TF-IDF vectorization, multiple machine learning models, model aggregation, real-time monitoring, and SMS alerts—act in concert to develop an effective phishing detection system. Below is a step-by-step description of the system flow according to the architectural design:

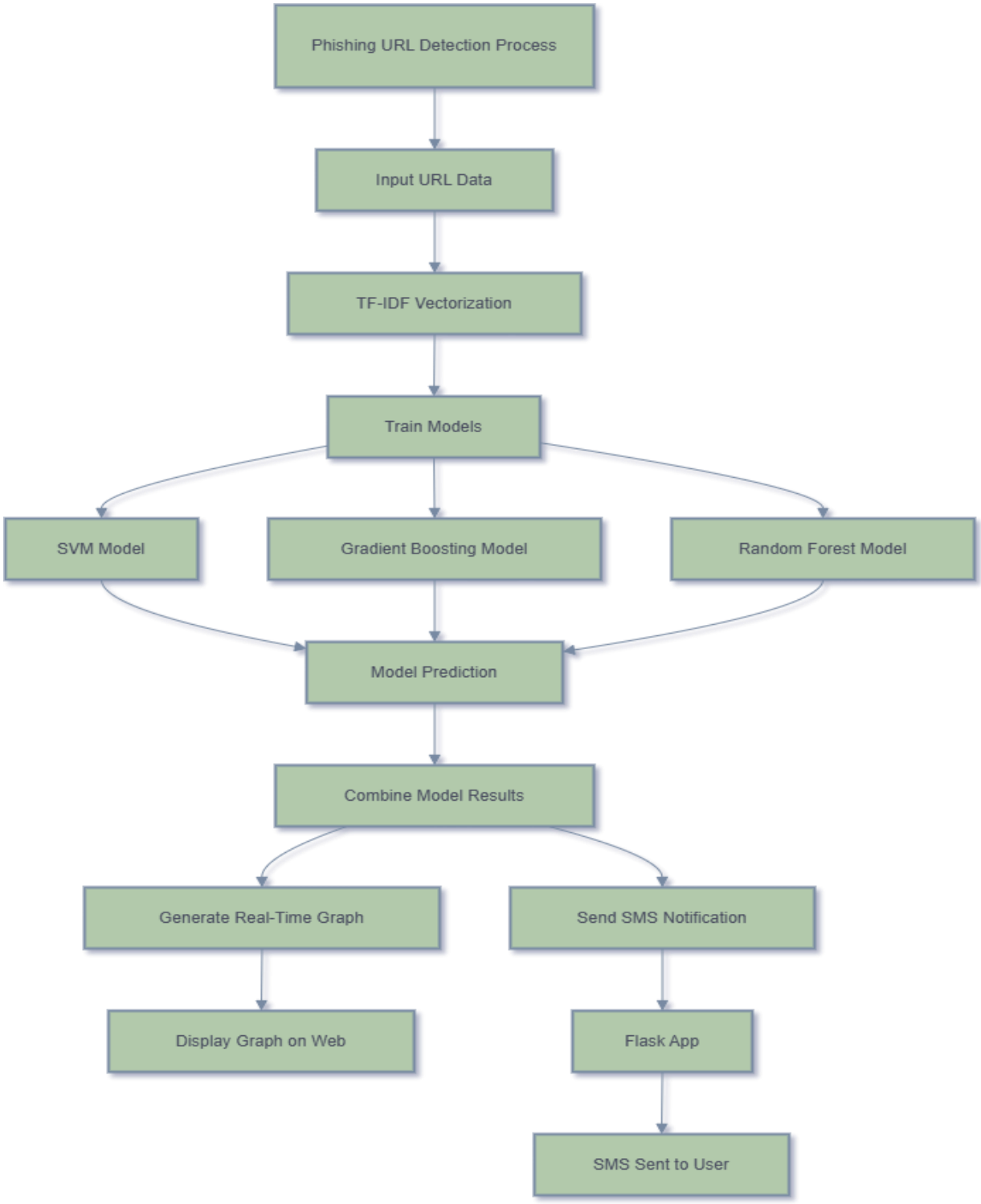

Figure. 3 Architectural Design

FIGURE 3: Demonstrates the system-level architecture of the Phishing URL Detection System. It depicts the data flow from URL input to feature extraction via the TF-IDF algorithm. The data is then processed through several machine learning models (SVM, Random Forest, Gradient Boosting) for classification purposes. The output from these classifiers are combined to arrive at a final decision, which is thereafter displayed on a web interface and alerted through SMS notifications. This architecture supports an end-to-end and real-time phishing detection process.

- **INPUT URL DATA:**

Users or automated systems introduce URLs into the detection system at the start of the process. The URLs may be from various things such as websites, or automated crawlers. The system can accommodate both single user inputs and batch inputs from other systems to ensure that it is flexible in its operation. Data Source: URLs may be entered directly by users, harvested from web traffic, or fetched from email links.Purpose: This step prepares the URLs to be ready for analysis during the later phases of the detection process.

- **FEATURE EXTRACTION:**

TFIDF vectorization transforms the URLs into a format which are relevant for the AI technique models. After the URLs have been gathered, they are passed through a feature extraction stage by the application of TF-IDF vectorization. This approach converts the raw text of the URLs to numerical vectors that can be analyzed by models. TF-IDF Transformation: URLs are segmented into into their individual terms, and TF-IDF assists in identifying the phishing URL and unique terms. This is achieved by employing the frequency of a term in a particular URL (TF) and its relevance in all URLs in the dataset (IDF). This step identifies suspicious words or patterns that may suggest phishing. Result: Following this conversion, the system gains the numerical features that represents each URL, which can be utilized by machine learning models in order to make predictions.

- **TRAINING PHASE:**

The second step entails training the machine learning models using a labeled dataset containing both phishing and normal URLs. The models are employed for (SVM), Gradient Boosting, and Random Forest. All three models are trained to identify the features that distinguish phishing URLs from legitimate ones.

- **SVM MODEL:**

SVM is particularly effective in binary classification problems, and it operates well in high-dimensional feature spaces, making it perfect for discriminating between phishing and legitimate URLs.

- **GRADIENT BOOSTING:**

The model improves iteratively by making the best predictions possible by combining results of many weak learners (models). It is able to learn complex relationships in the data, so it is good at detecting subtle phishing patterns.

- **RANDOM FOREST:**

Random Forest employs an ensemble trees, each of which are trained on a random subset of the data. By combining their outputs, RF provides a strong and consistent classification, reducing over-fitting.

- **PREDICTION:**

The learned models automatically classify if the URLs are phishing or not. When the models are trained, the system is URLs. The models examine the numerical features from the TF-IDF conversion and label each URL as phishing or legitimate. Model Independence: Each of the three models (SVM, Gradient Boosting, and Random Forest) makes its own individual classification independently. prediction, depending on its learning and the extracted features from the URL. The result is a collection of predictions, with each model giving a label (phishing or legitimate) to the input URL.

- **AGGREGATION:**

The models' predictions are aggregated to make a final decision using methods such as majority voting or weighted averaging. The system's output is more trustworthy by taking advantage of the strengths of several models. Ensemble Techniques Aggregation may be done using various techniques: Majority Voting: In the event of most models voting a URL as phishing, it is classified as phishing.Weighted Averaging: Where different importance levels are assigned to models, the models' predictions are weighted. The decision is made based on the weighted average.Stacking: A secondary learning process that is applied to aggregate the output of the models for sophisticated decision.

- **REALTIME MONITORING:**

The system provides updated real-time graphs and visualizations, such that users can track the phishing detection process. The system updates and continuously tracks the progress of the phishing detection in real-time. This includes the generation of dynamic graphs and visualizations reflecting the system's detection performance. Graphical Output: Key metrics, including how many number of detected phishing URLs, detection trends through time, and performance measures (accuracy, precision, recall) are presented in real-time graphs. This allows system administrators to track the detection process in real time.Visualization: New data processed updates the graphs, giving a current picture of phishing detection activity.Real-time monitoring prevents the detection system from being idle and guarantees that any new trends or emergence threats are rapidly detected and examined.

- **ALERTING:**

When a phishing URL is identified, the system sends an alert to inform users of the threat. The alerting mechanism may take necessary action to prevent or report phishing. SMS Notification: An automated message through SMS is sent to the user whenever a phishing URL is identified. This is often carried out using a backend program such as Flask, which controls the process of notification. User Awareness: The SMS notification notifies the user of the phishing threat detected, enabling them to take immediate action like avoiding the link or reporting the relevant authorities. This process is essential to phishing attacks and enables proactive actions to be taken in response to threats. This design of architecture and algorithms ensures that the Phishing URL Detection System is precise, scalable, and proactive in identifying phishing threats. Every module works together to create an integrated, intelligent defense system against phishing attacks.

IV. CONCLUSION, FUTURE WORK AND APPLICATION:

A. CONCLUSION

By way of conclusion, the Phishing URL Detection System is a major innovation in the war on phishing attacks, offering a very accurate, scalable, and adaptive solution for detecting malicious URLs. Phishing is still one of the most common and destructive cyber threats, and classical detection approaches tend to fail in handling the dynamism and complexity of phishing methods. Based on sophisticated machine learning algorithms such as SVM, Gradient Boosting, and Random Forest, and TFIDF vectorization for feature extraction, this system presents a sound method of phishing detection that can continuously improve and adapt itself to emerging threats.The use of ensemble learning methods provides assurance that the system provides more accurate predictions through the aggregation of different models' strengths, hence less risk of mistakes such as false negatives and positives. This aspect is especially valuable considering the growing sophistication of phishing attacks that tend to imitate valid URLs or use new ways to evade detection.The capability of the system to offer real-time monitoring via graphical visualizations and phishing trend tracking provides a vital situational awareness layer for users and security teams. The SMS alert mechanism guarantees that users are immediately notified when a phishing URL is identified, allowing for quicker response and enhanced security. Such proactive alert mechanisms are essential in countering the effects of phishing attacks and enhancing overall cybersecurity defenses. In addition, the scalability and flexibility of the system allow it to be deployable across industries and environments, ranging from individual users to large enterprises. The system's functionality is continually perfected through the addition of new information, making it effective despite the development of phishing techniques. The Phishing URL Detection System is a powerful and smart solution to the increasing menace of phishing attacks. With the use of multiple machine learning algorithms, sophisticated feature extraction processes, real-time forecasting, and SMS notification processes, the system promises accuracy and responsiveness. Every module, from data gathering to notification, functions together in a concerted effort to ensure that phishing threats are detected and responded to in a timely manner. Having a modular design makes it possible to maintain easily, scale, and make improvements constantly based on new information. this system contributes significantly toward enhancing cyber security and protecting users from malicious online activities.

B. FUTURE WORK:

The Phishing URL Detection System can be further improved in a number of important areas to make it more effective and flexible. As phishing methods change, so must the systems used to detect them. The following are a number of areas for future development that could greatly improve the system's functionality:RealTime Protection: To offer more immediate protection against phishing attacks, the system might be coupled with real-time software such as browsers or email clients. This integration would enable the system to constantly scan the URLs accessed by users while surfing online or opening emails. By marking phishing attempts promptly as

users surf online, the system would be able to block malicious sites ahead of users before they can be tricked. This anticipatory protection would provide an extra useful layer of security, preventing users from accessing phishing sites in the first place, instead than acting afterward. Larger Dataset: One of the most important enhancements that can make the system more accurate is increasing the dataset upon which the machine learning model is trained. At present, the system is designed to detect conventional phishing URLs, but phishing is a multifaceted attack and occurs in numerous forms. For example, voice phishing (vishing) and social media phishing are increasingly becoming issues, as the attackers now target victims via phone calls and social networks. By through incorporating more diverse phishing examples in the dataset, training the system can recognize and block more types of phishing attacks, its strength and variability in countering new forms of attacks enhance. Deep Learning Models: Although today's system operates using algorithms like SVM, Gradient Boosting, and Random Forest, integrating deep models like neural networks into it could make a big difference improvements.Deep learning methods are best at detecting intricate patterns in vast amounts of data, which may be especially helpful in finding more advanced and subtle phishing attempts that conventional algorithms may overlook. Through the use of deep learning, the system would be able to examine a greater number of URL features, learn from massive datasets, and detect even the most advanced phishing patterns with greater accuracy. Better Adaptability: To be effective despite the ever-changing phishing methods, the capacity of the system to learn automatically from new phishing data needs to be enhanced. At present, the system needs to be updated manually with new datasets and retrained models. However, integrating a mechanism for continuous learning-where the system automatically adapts to new phishing patterns without human intervention-would allow the system to stay Up To Date with emerging Threats This flexibility would allow the system to recognize newly discovered phishing tactics without constant human intervention. User Behavior Monitoring: Another potential direction for future work is taking advantage of user behavior monitoring to improve phishing detection. By examining user behavior, including browsing history, email usage, and interaction with links, the system might identify patterns indicating possible phishing attacks even before they occur. For instance, if one keeps visiting trustworthy banking websites on a regular basis, but somehow suddenly gets email notification containing a phony link disguised as a website of a bank, the system would mark this as abnormal behavior.Introducing behavioral analysis, the system could offer an added layer of security by being able to prevent phishing attacks by foreseeing threats and warning the users before being a victim.

C. APPLICATION:

The Phishing URL Detection System is of very broad practical application across many fields of cybersecurity. It can be built into web browsers to automatically identify or block phishing URLs prior to their being visited by users, improving internet safety. In email clients, it can be applied to scan for embedded links in real-time and flag suspicious mail, cutting down on risk of phishing attacks on unsuspecting users. Enterprises can deploy this system within their internal

network security infrastructure to actively shield employees from falling prey to phishing scams, particularly in sectors such as banking, healthcare, and e commerce where sensitive information is regularly dealt with. Educational institutions can also use this system to raise awareness among students and staff regarding safe browsing practices and identifying dangerous connections. It may also be implemented in mobile and web apps to offer real-time warnings and help make the internet a safer place. Governments and cyber-security organizations can take advantage of it being used in national threat intelligence systems to enable bulk detection and blocking. With its scalable and modular architecture, the platform is very flexible to the changing phishing methodologies, rendering it a potent tool in contemporary digital security measures.

V. ANNEXURE:

Architecture Diagram Breakdown for Phishing URL Detection System:

the design of the Phishing URL Detection System consists of multiple interconnected modules, each with a specific function within the overall detection pipeline. It starts with the Data Collection Module, which collects URLs from a variety of sources like web crawlers, emails, user inputs, and external threat databases. The raw data collected is then fed into the Data Preprocessing Module, where it is cleaned, normalized, and tokenized for subsequent analysis. Feature Extraction Module breaks down the URLs to identify important features, including domain name features, URL length, and special characters, employing techniques such as TF-IDF vectorization. The key features are utilized in the Model Training Module, where machine learning models like SVM, Random Forest, and Gradient Boosting are utilized to train a model that can classify phishing Legitimate URLs. After training, the Prediction Module applies these models to categorize incoming URLs in real-time. These results are presented in the Visualization and Reporting Module, which creates graphs and interactive dashboards for tracking the detection process. If a phishing URL is found, the Notification Module sends alerts through SMS or email to inform users or administrators of the threat. Finally, the System Maintenance and Updates Module keeps the system current with emerging phishing methods by periodically retraining models and fixing any system problems. This modular design provides a strong, scalable, and effective phishing detection system. The Phishing URL Detection System has numerous real-world applications in different areas of cybersecurity. It can be incorporated into web browsers to automatically warn or block malicious URLs prior to users access them, to increase online safety. In e-mail clients, it can also be employed for scanning enabled links in real-time and flagging dangerous messages, inhibiting phishing scams on unlnerable users. Companies can set up this system in their own internal network security components to actively keep employees safe against phishing scams, particularly in business models such as banks, hospitals, and online shoppers where sensitive information is often touched. Educational institutions may also employ this system to raise awareness among students and teachers about safe browsing and identifying malicious links. Moreover, it can be implemented in web and mobile applications to give live alerts, adding to a more secure online world.

VI. REFERENCES:

[1] N. Z. Harun, N. Jaffar, and P. S. J. Kassim, ‘‘Physical attributes significant in preserving the social sustainability of the traditional Malay settlement,’’ in Reframing the Vernacular: Politics, Semiotics, and Representation. Springer, 2020, pp. 225238.

[2] S. Bell and P. Komisarczuk, ‘‘An analysis of phishing blacklists: Google safe browsing, OpenPhish, and PhishTank,’’ in Proc. Australas. Comput. Sci. Week Multiconf. (ACSW), Melbourne, VIC, Australia. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–11, Art. no. 3, doi: 10.1145/3373017.3373020.

[3] A. K. Murthy and Suresha, ‘‘XML URL classification based on their semantic structure orientation for web mining applications,’’ Proc. Comput. Sci., vol. 46, pp. 143–150, Jan. 2015.

[4] R. Prasad and V. Rohokale, ‘‘Cyber threats and attack overview,’’ in Cyber Security: The Lifeline of Information and Communication Technology. Cham, Switzerland: Springer, 2020, pp. 15–31.

[5] T. Nathezhtha, D. Sangeetha, and V. Vaidehi, ‘‘WCPAD: Web crawling based phishing attack detection,’’ in Proc. Int. Carnahan Conf. Secur. Technol. (ICCST), Oct. 2019, pp. 1–6.

[6] M. Zouina and B. Outtaj, ‘‘A novel lightweight URL phishing detection system using SVM and similarity index,’’ Hum.Centric Comput. Inf. Sci., vol. 7, no. 1, p. 17, Jun. 2017.

[7] D. M. Divakaran and A. Oest, ‘‘Phishing detection leveraging machine learning and deep learning: A review,’’ 2022, arXiv:2205.07411.

[8] A. Akanchha, ‘‘Exploring a robust machine learning classifier for detecting phishing domains using SSL certificates,’’ Fac. Comput. Sci., Dalhousie Univ., Halifax, NS, Canada, Tech. Rep. 10222/78875, 2020.

[9] H. Shahriar and S. Nimmagadda, ‘‘Network intrusion detection for TCP/IP packets with machine learning techniques,’’ in Machine Intelligence and Big Data Analytics for Cybersecurity Applications. Cham, Switzerland: Springer, 2020, pp. 231–247.

[10] M. Kline, E. Oakes, and P. Barford, ‘‘A URLbased analysis of WWW structure and dynamics,’’ in Proc. Netw. Traffic Meas. Anal. Conf. (TMA), Jun. 2019, p. 800.

[11] A. A. Ubing, S. Kamilia, A. Abdullah, N. Jhanjhi, and M. Supramaniam, ‘‘Phishing website detection: An improved accuracy through feature selection and ensemble learning,’’ Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 1, pp. 252–257, 2019. Dept. of CSE 55 2024-25 A MULTIALGORITHMIC APPROACH FOR PHISHING DETECTION: LEVERAGING TFIDF VECTORIZATION WITH SVM, GB, AND RF