International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Detecting Phishing Links Analysis Using Machine Learning

K.N.S.B.V. Manjusha[1], Dr. D. Jaya Kumari[2]

[1]Student, Department of Computer Science, Sri Vasavi Engineering College (A), Pedatadepalli, Tadepalligudem - 534101.
[2]Professor & Head of the Department, Department of CSE, Sri Vasavi Engineering College (A), Pedatadepalli, Tadepalligudem - 534101.

Abstract
This proposed strategy for identifying phishing websites employed the Gradient Boosting Classifier model, focusing on various aspects of URL significance. By meticulously extracting and comparing different characteristics between legitimate and phishing URLs, our approach leverages the Gradient Boosting Classifier to identify phishing URLs effectively. The study's findings underscore the successful application of our suggested approach in real-time, demonstrating its ability to distinguish between legitimate and bogus websites. Given the relentless evolution of phishing techniques facilitated by advancing technology, employing anti-phishing methods is imperative. Phishing attacks, which often rely on deceptive websites closely resembling genuine ones in appearance and language, pose a significant threat. Machine learning emerges as a robust tool in thwarting such assaults, offering the ability to discern subtle patterns indicative of malicious intent. Phishing remains a preferred tactic for attackers due to its effectiveness in bypassing traditional security measures. By duping unsuspecting users into clicking seemingly authentic yet malicious links, attackers exploit human vulnerability, highlighting the importance of proactive detection mechanisms. In this context, our utilization of the Gradient Boosting Classifier underscores the efficacy of machine learning in fortifying defences against phishing attacks. As cyber threats evolve, embracing innovative approaches like machine learning becomes essential in safeguarding against emerging risks.

Keywords: Phishing attacks, Machine Learning, Gradient Boost Classifier, URL Features.

1. INTRODUCTION

The primary goal of a learner is to extrapolate knowledge from their prior experiences. In this context, generalization refers to the capability of a learning system to accurately handle new and unseen tasks or examples following exposure to a training dataset. These training instances are sourced from a generally unknown probability distribution, representing the broader spectrum of occurrences. The learner's objective is to construct a generalized model of this distribution, enabling it to make sufficiently accurate predictions for novel cases. The computational examination of machine learning algorithms and their efficacy falls under computational learning theory, a field within theoretical computer science. Beyond simply evaluating performance, learning theorists delve into aspects such as time complexity and the feasibility of learning processes. Within computational learning theory, feasibility is assessed based on whether a computation can be executed in polynomial time. Time complexity findings in this domain

generally fall into two categories: positive results demonstrate that certain classes of functions can indeed be learned within polynomial time.

1.1 Gradient Boosting Algorithm

The fundamental concept driving this algorithm is sequential model building, where each subsequent model endeavours to minimize the errors made by its predecessor. The question arises: how exactly is this achieved? How do we go about diminishing these errors? The solution lies in constructing a new model based on the errors, or residuals, of the preceding model.

Gradient Boosting is a robust algorithm that can detect subtle URL patterns that indicate malicious intent regarding machine learning-based phishing link detection. An ensemble learning method called gradient boosting builds a robust prediction model by progressively combining several weak learners, usually decision trees. This iterative process gradually improves the overall predictive accuracy by fitting new models to the residuals of earlier models. Specifically, Gradient Boosting can handle imbalanced data frequently present in phishing detection datasets because it ensures balanced learning by giving higher weights to incorrectly classified instances of the minority class during training. Researchers can also determine which URL features, like domain patterns or structural anomalies, are most suggestive of phishing activity thanks to the algorithm's insights into feature importance. The model's ability to withstand overfitting, attained by shrinkage and tree pruning, guarantees that it applies well to new data, essential for adjusting to phishing tactics that change over time. Gradient Boosting reduces false positives with its high predictive accuracy, essential for trustworthy phishing detection. Additionally, because of its adaptability, the model can incorporate different features extracted from URLs, such as linguistic, domain-based, and structural features, improving its ability to distinguish between phishing and legitimate URLs. In conclusion, Gradient Boosting is essential to building robust and accurate machine learning models for phishing link detection, enabling enterprises to counteract the constantly changing threat landscape of phishing attacks proactively.

2. LITERATURE SURVEY

The cornerstone of the software development process lies in conducting a comprehensive literature review. This crucial step involves delving into existing research conducted by numerous authors relevant to our work. By carefully analyzing and synthesizing key articles, we gain valuable insights that serve as the foundation for extending our work further. Through this process, we acknowledge the contributions of previous studies and leverage their findings to inform and enrich our endeavours.

Karin, Abdul, et al. [1] examined the efficacy of gradient boosting classifiers (GBM) as a machine learning model for detecting phishing URLs. Their research contributes to the field by offering a comprehensive assessment of privacy and security decision-making literature spanning multiple disciplines. Specifically, the study concentrates on research to aid individuals in making informed privacy and security decisions, employing soft paternalistic interventions to guide users towards more advantageous choices subtly. Through a thorough analysis, the article explores the potential benefits of these interventions and sheds light on their limitations. Moreover, it identifies crucial ethical, design, and research challenges in implementing such interventions. This multi-faceted investigation underscores the complex interplay between technology, human behaviour, and ethical considerations in privacy and security decision-making.

Nowroozi et al. [2] specifically delve into detecting phishing websites using URLs and gradient-boosting techniques. They address a critical aspect of cybersecurity, focusing on the susceptibility of individuals to

social engineering attacks due to their inherent trust in others and their tendency to disclose personal information readily. The study investigates the efficacy of two interventions to safeguard users against such attacks: priming through cues designed to heighten awareness about the risks associated with social engineering cyber-attacks and warnings cautioning against divulging personal information. Conducting their research within the shopping district of a medium-sized town in the Netherlands, the study scrutinizes the behaviour of visitors in response to these interventions. Alarmingly high disclosure rates are revealed, with 79.1% of subjects providing their e-mail addresses and 43.5% disclosing bank account information.

Furthermore, among online shoppers, an overwhelming majority—89.8%—reveal the type of product(s) they purchased, while 91.4% disclose the name of the online shop where these purchases were made. These findings underscore the pressing need for effective measures to combat social engineering attacks and protect individuals' sensitive information. By shedding light on the extent of vulnerability exhibited by individuals in real-world scenarios, the study highlights the urgency of implementing robust cybersecurity strategies and user awareness campaigns. Moreover, it emphasizes the importance of ongoing research and innovation in developing sophisticated detection mechanisms, such as gradient-boosting algorithms, to identify and mitigate the risks of phishing websites and other cyber threats.

M. El-Ally, El-Sayed M, Lichen, et al. [5] diverge from gradient boosting in detecting phishing websites through URLs. Instead, it focuses on a deep learning-based framework for detecting and visualising malicious online advertisements. The innovative approach presented in this paper revolves around using probabilistic neural networks (PNNs) to detect phishing websites. Additionally, the authors explore the integration of PNNs with K-medoid clustering to reduce complexity while maintaining detection accuracy significantly. Through experimental validation, the results reveal the feasibility of constructing more precise models. Remarkably, even with a reduction in complexity exceeding 40%, the approach achieves over 97% accuracy, with minimal false errors. By introducing this novel methodology, the paper advances phishing detection techniques and underscores the potential of leveraging probabilistic neural networks in cybersecurity applications. The demonstrated effectiveness of the integrated approach highlights its practical viability in real-world settings, emphasizing the importance of exploring diverse methodologies to address evolving cyber threats. Moreover, the findings underscore the significance of optimizing model complexity to balance computational efficiency and detection accuracy, offering valuable insights for future research endeavours in cybersecurity.

Ch., Rupa, et al. [11], led by Cheng Huang, offer a unique perspective on safeguarding users against identity theft attacks by detecting malicious links through host-based and linguistic features of URLs. Despite not explicitly utilizing gradient boosting for phishing website detection, the research delves into a crucial aspect of cybersecurity. The paper introduces a methodology that detects "domain fluxes" within DNS (Domain Name System) traffic. Specifically, the researchers target patterns inherent to domain names generated algorithmically, distinguishing them from those created by human users. By analyzing the distribution of alphanumeric characters and bigrams across domains mapped to the same set of IP addresses, the study aims to uncover suspicious activities indicative of potential malicious intent. This innovative approach underscores the importance of leveraging sophisticated techniques to identify and mitigate emerging cyber threats, such as domain fluxes associated with identity theft attacks. By scrutinizing the underlying patterns within DNS traffic, the research offers valuable insights into the behaviour of malicious actors and their tactics for evading detection.

Furthermore, the methodology developed in this study presents a proactive means of enhancing

cybersecurity defences, enabling organizations to preemptively address potential vulnerabilities and protect users from identity theft and other forms of cyber exploitation. Overall, the research conducted by digital landscape.

3. EXISTING SYSTEM

Many noteworthy methods have been put out in the field of machine learning-based phishing link detection. Huang et al. (2009) presented frameworks that use URL token analysis to predict phishing pages that frequently imitate the CSS style of their target pages, allowing them to differentiate between phishing and legitimate websites based on page section similarity. Their methodology consists of dissecting the URL tokens and contrasting them with the URL tokens of the target page, emphasizing the similarity of page sections. By distinguishing minute variations between authentic and fraudulent web pages through their URL architecture and CSS styles, this technique seeks to achieve high prediction accuracy.

Marchal et al. (2017) presented a method based on analyzing accurate site server log data that distinguishes phishing websites. Their method looks for trends and anomalies in server log data that point to phishing activity. Based on this methodology, the Off-the- Hook program provides several benefits: high accuracy, total independence, flexibility in terms of language, quick decision-making, and resilience against dynamic phishing assaults. Because of the technique's capacity to examine server log data, it is possible to identify phishing websites in real time and take preventative action against people visiting these dangerous websites.

Ayden et al. presented a classification system that uses subset-based feature selection techniques and extracts URL features for phishing website identification. To detect phishing websites, their method relies on extracting information from URLs, such as domain name length, the presence of memorable characters, and URL length. Their objective is to enhance the precision and effectiveness of their classification algorithm by identifying the most pertinent elements for phishing website detection through subset- based feature selection techniques. In the existing system, machine learning techniques such as logistic regression, multinomial naive Bayes, and XGBoost are evaluated; logistic regression shows better performance. A well-liked machine learning approach for binary classification problems, logistic regression is ideal for identifying phishing websites. The suggested method tokenizes words and applies stemming to the model as a preprocessing step. Data processing is done to encode or convert data so that machines may quickly transfer it. The entire comparison shows that logistic regression achieved an accuracy of

96.63 per cent.

Limitations of the Existing System

1. Lack of a dedicated user interface: The current systems must provide a specific user interface, which may hinder user interaction and ease of use.
2. Limited to binary outcomes: The existing model is designed to predict dichotomous outcomes and cannot handle continuous variables, restricting its applicability in scenarios requiring continuous predictions.
3. Impact of small sample sizes: The accuracy of the existing model may be compromised when the sample size is too small, leading to less reliable predictions.
4. Risk of overfitting: Overfitting in the existing model occurs when the model learns the noise in the training data rather than the underlying pattern, potentially affecting its ability to generalize to new

data.

4. PROPOSED SYSTEM

The project's primary development platform will be a website with an accessible, user- friendly interface that is interactive for all users. This website is a vital tool for determining whether a website is legitimate or fraudulent. Web development languages used in its creation include HTML, CSS, JavaScript, and Python's Flask framework. CSS is primarily used to add effects to ensure a user-friendly experience and enhance the visual beauty of the website. Considering the diverse user base, the website has been meticulously crafted to ensure ease of use for all users.

The proposed system is trained using a carefully curated dataset that comprises various features relevant to determining the legitimacy of a website. The dataset does not include website URLs, focusing on other vital features. These features are essential for the system to effectively differentiate between legitimate and phishing websites. To achieve this, the Gradient Boosting Classifier is employed. This classifier is trained using the dataset, enabling it to learn patterns and characteristics that distinguish phishing websites from legitimate ones. Once the training is complete, the classifier can evaluate a URL accurately. The system promptly alerts the user if the website is identified as phishing. Conversely, if the website is legitimate, the user receives confirmation. The Gradient Boosting Classifier demonstrates an impressive accuracy rate of 97% in detecting phishing websites.

4.1 Advantages of the Proposed System

1. **User-Friendly Interface**: The project includes an intuitive and easy-to-use user interface, ensuring users can navigate the system without encountering difficulties.
2. **Comprehensive Feature Set**: The model is trained using a wide range of features, enabling it to effectively differentiate between legitimate and phishing websites based on various characteristics.
3. **High Accuracy**: The system accurately identifies phishing websites, providing users with reliable and trustworthy results.
4. **Superior Performance**: Compared to other models, the proposed system demonstrates higher accuracy rates, making it a more effective tool for detecting phishing attempts.
5. **Faster Training**: The proposed system can train more quickly, especially when dealing with larger datasets, allowing for faster deployment and implementation.
6. **Support for Categorical Features**: Most versions of the proposed system offer support for handling categorical features, enhancing the model's ability to process and analyze diverse data types.
7. **Handling of Missing Values**: Some versions of the proposed system are equipped to handle missing values natively, ensuring the model can effectively process incomplete datasets without compromising accuracy.

5. Dataset Description

Kaggle [22] collected the phishing URL detection dataset and included several variables to differentiate between phishing and authentic websites. Some of these qualities include symbols like "@" in the URL, its length, whether it employs an IP address, and whether it uses "//" to redirect users to another page. The collection also contains attributes about the URL's structure, like the number of subdomains, the presence of prefixes or suffixes, and if the URL uses HTTPS. Additional characteristics include domain registration duration, non-standard ports' usage, and a favicon's existence. In addition, the dataset

contains characteristics of the behaviour and content of the website, like the usage of server form handlers, the existence of aberrant URLs, and the inclusion of links in script tags.

Furthermore included are features associated with website functionality, like website forwarding, customizable status bars, and the ability to turn off right-click capability. The dataset contains information on Google indexing, DNS recording, website traffic, and domain age. Lastly, 'class', a target variable in the dataset, determines if a URL is considered legitimate or phishing. To create and hone machine learning models for phishing URL detection, this dataset offers an extensive feature set.

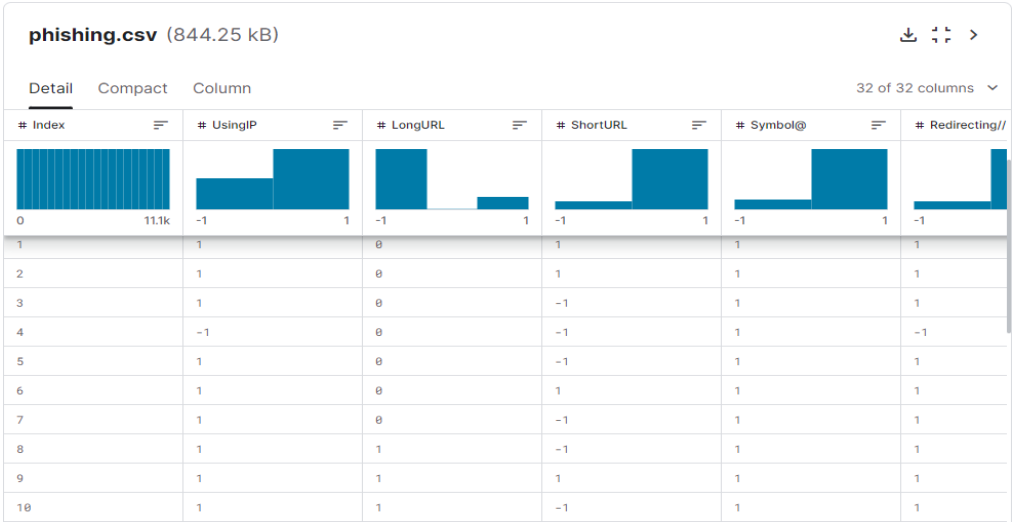

Figure 1 Preview of the Dataset

6. Methodology

Figure 2 illustrates the process of acquiring the initial dataset from Kaggle, which comprises features associated with URL properties. The suggested methodology for phishing URL detection includes multiple essential steps. The input dataset comprises features associated with URL properties and is first acquired via Kaggle.

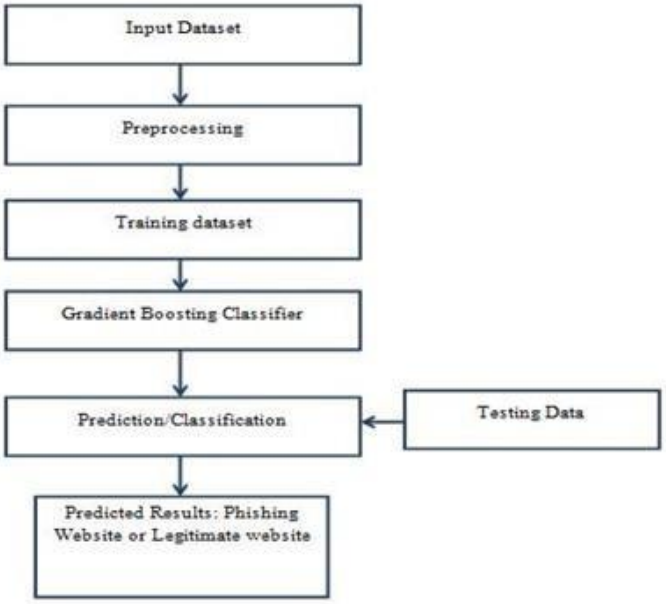

Figure 2 Proposed Methodology

Preprocessing is done to clean the dataset and prepare it for training. This involves scaling numerical features, encoding categorical variables, and managing missing values. Next, training and testing sets are created from the preprocessed dataset. A robust machine learning technique called Gradient Boosting Classifier, which can handle complicated datasets and generate accurate predictions, is trained on the training dataset. Upon training, the model is applied to the testing dataset to provide predictions. Lastly, a web application with an intuitive user interface for interacting with the system and viewing the URL classification results displays the projected outcomes. This methodology offers an easy-to-use interface for obtaining the findings and successfully detects phishing URLs using machine learning techniques.

6. System Architecture

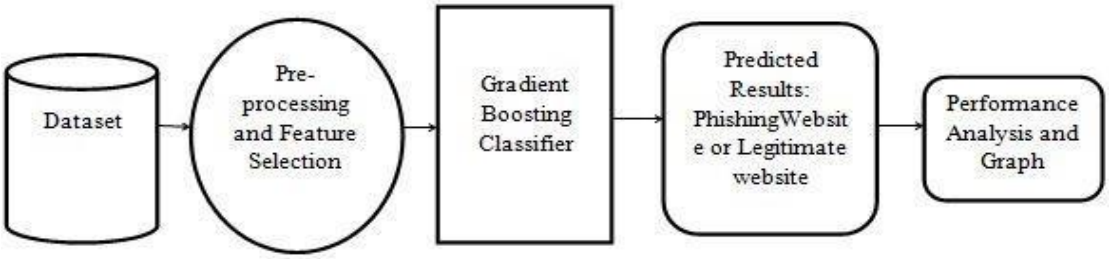

Figure 3 System Architecture

Figure 3 elaborates on the system architecture for phishing URL detection. It includes multiple features associated with URL properties. This dataset undergoes feature selection and preprocessing steps to clean the data and extract relevant features for the model's training. The preprocessed data is then sent into the Gradient Boosting Classifier, a highly effective machine learning technique for handling complex datasets. The classifier predicts if a particular URL will link to a legitimate website or a phishing one based on the patterns it has learned from the training set of data. Performance metrics such as recall, accuracy, precision, and F1 score are then compared to the anticipated results. These metrics are employed to assess how well the algorithm detects phishing URLs. Graphical representations of the performance measures are also generated to give a visual sense of the model's performance. This system design uses machine learning techniques and performance analysis to identify phishing URLs and assess the model's performance. A Data Flow Diagram (DFD), a bubble chart, is the graphical representation of a system's components and the data flow between them. It consists of processes representing the system's functions, data stores holding data for the processes, and external entities interacting with the system. The DFD shows how information moves through the system and undergoes transformations, depicting data flow from input to output. It can represent a system at different levels of abstraction and can be partitioned into levels to represent increasing information flow and functional detail. Overall, the DFD is a valuable
modelling tool for understanding and visualizing how data moves through a system and its transformations.

7. Experimental Results

Numerous studies have demonstrated that, depending on several variables, phishing link detection model accuracy varies from 85% to 99%. These variables include the complexity of the features used, the quantity and calibre of the dataset, and the advanced machine learning methods used. Because ensemble methods can handle complicated feature interactions well, they have shown to be very useful in this

setting. Examples of these methods are random forests. However, the effectiveness of these models can differ depending on several variables, such as the phishing attack types targeted (e.g., spear phishing, SMS phishing, and e-mail phishing), the variety of attributes taken into account, and the model's capacity to adjust to changing phishing tactics.

In this study, a phishing detection model was developed using the Gradient Boosting Classifier, achieving an impressive accuracy rate of 98.9%. The Gradient Boosting Classifier is known for its ability to handle complex datasets and produce high-quality predictions, making it a suitable choice. The high accuracy of the model indicates its effectiveness in distinguishing between legitimate and phishing URLs, highlighting its potential for enhancing cybersecurity measures. Additionally, a web application was implemented as part of the study to showcase the model's functionality and provide users with a practical tool for detecting phishing URLs.

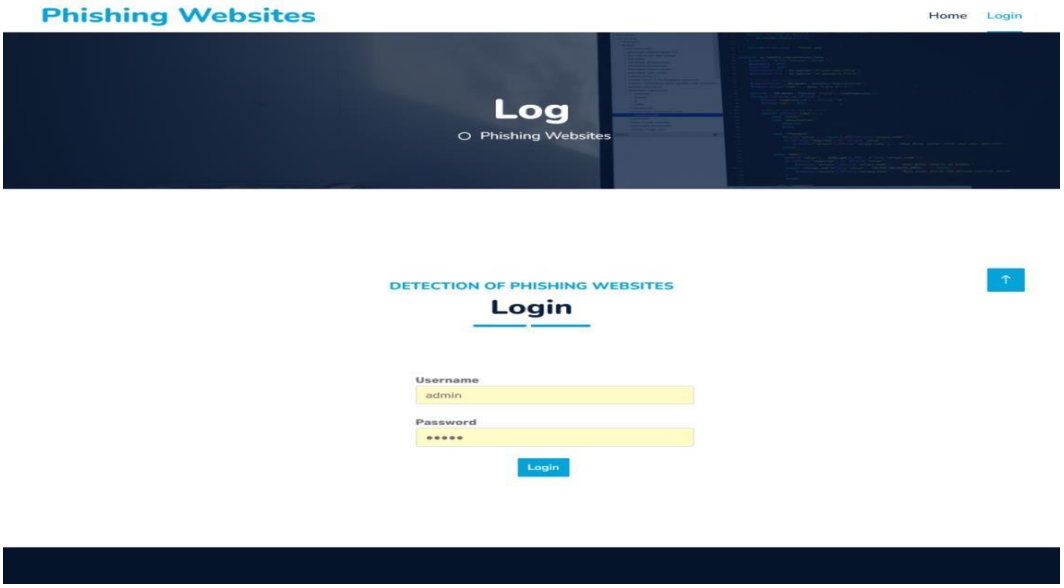

Figure 4 Application Login Page

The first page of the web application for phishing URL detection is the login page, as shown in Figure 4. The user must log in using their username and password. Following the successful login, the user is taken to the home page.

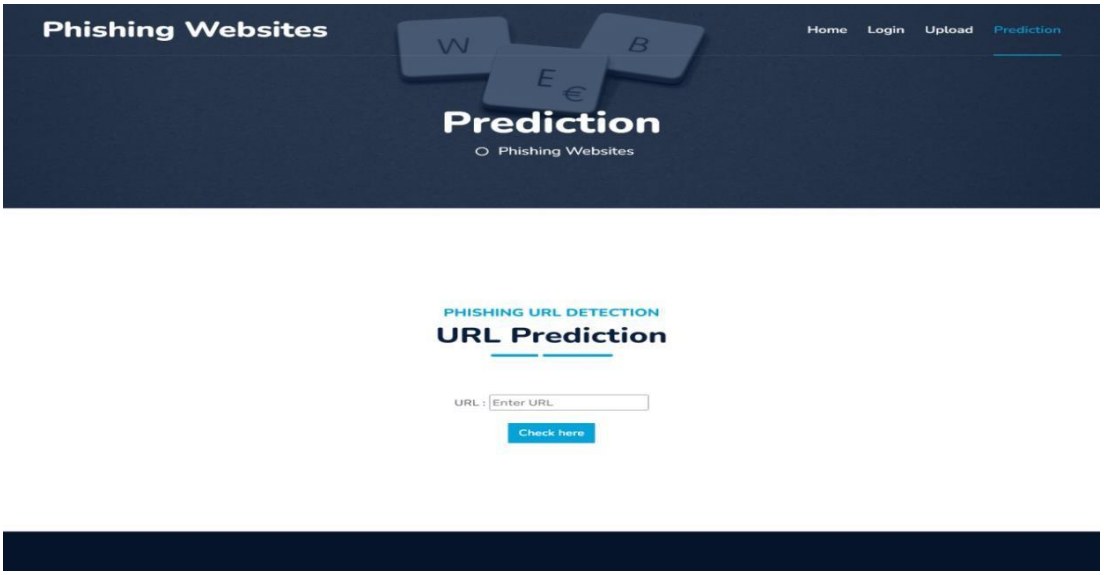

Figure 5 Application Home Page

Once the user logs in successfully, the home page shown in Figure 5 is displayed to the user. The page contains the title and a field for entering the URL.

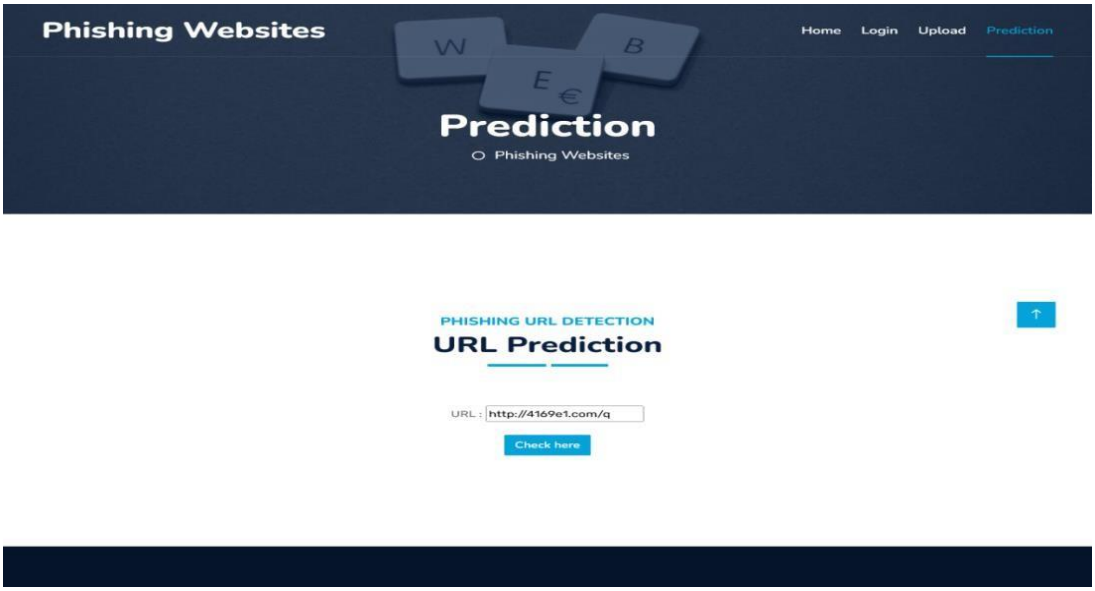

Figure 6 URL Input

As shown in Figure 6, the user has to input the URL to identify whether it is legitimate or phishing. After entering the URL, the user must click the "Check here" button to make the predictions.

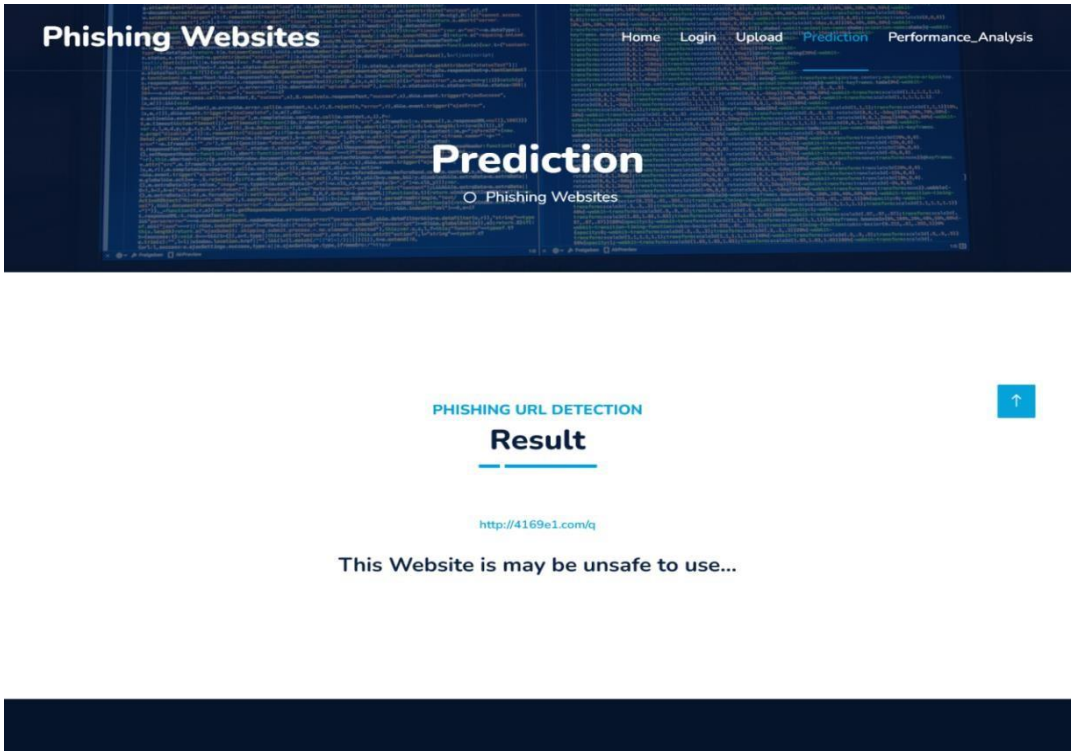

Figure 7 Prediction and Output

Once the user clicks on the "Check here" button, the user is redirected to the Results page, as shown in Fig 7. Here, the actual prediction happens. The URL entered is passed to the machine learning model running in the application. The model takes the input and predicts if the URL is Phishing or Legitimate, and finally, the result is displayed on the screen. Here, the result is shown to be "This Website may be unsafe to use", meaning that the entered URL is a phishing URL.

Figure 8 Performance Analysis

Fig 8 shows the model's performance analysis. The Precision, Recall, and F1 scores, along with the confusion matrix, are displayed.

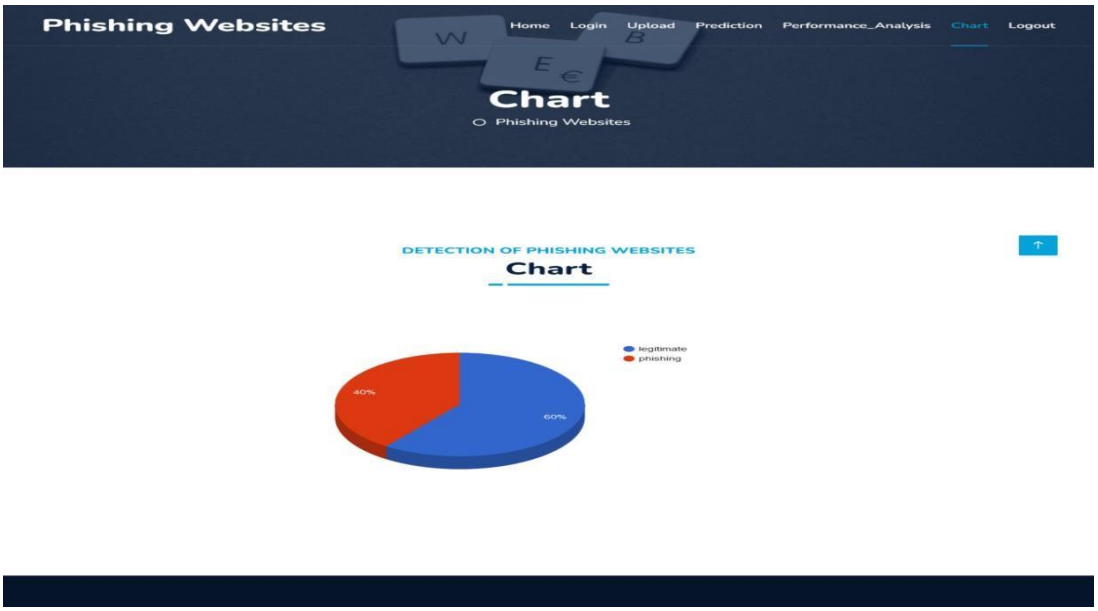

Figure 9 Distribution of Legitimate and Phishing URLs

Fig 9 shows a Pie chart showing the distribution of Phishing and Legitimate URLs. It can be observed from the graph that 60% of the URLs are Legitimate, and the remaining 40% belong to the Phishing URL category.

9. CONCLUSION

The detection of phishing links is a significant challenge in the field of cyber security because successful attacks can have disastrous consequences. Using machine learning methods, specifically the Gradient Boosting algorithm, is a big step in strengthening defences against these attacks. We have achieved remarkable success with our investigation into machine learning-based phishing link detection, explicitly

focusing on the Gradient Boosting algorithm. We have attained an astounding accuracy rate of 97.6%. This study shows how machine learning can effectively identify minute patterns in URLs that point to malicious intent. Through a rigorous feature extraction and analysis process involving structural, domain-based, and linguistic attributes, our model can discriminate between phishing and legitimate URLs. The Gradient Boosting algorithm's ability to support the higher accuracy attained highlights its effectiveness in precisely identifying phishing links. The high accuracy rate we achieved in our study is evidence of how resilient and flexible machine learning algorithms are in the face of changing cyber threats. Organizations can preventatively reduce the risks associated with phishing attacks by automating the detection process and utilizing sophisticated algorithms such as Gradient Boosting. This will protect confidential data and improve cybersecurity posture. There is a great deal of promise for future machine learning-based phishing detection research and development. Developing novel strategies, improving feature extraction processes, and fortifying model resilience will be essential to fending off increasingly cunning cybercriminals. Keeping efficacy in dynamic cyber landscapes will also require integrating real-time data feeds and regularly updating models to reflect new threats. In this machine learning model, a web application is developed.

10. FUTURE SCOPE

In a future work

Phishing detection systems are poised to evolve significantly, transitioning towards real-time monitoring capabilities for web traffic and user interactions. Through the continuous analysis of incoming data streams, these systems will be adept at swiftly recognizing and addressing emerging phishing threats. This proactive approach enables timely detection and response, minimizing the potential impact of phishing attacks on individuals and organizations. By harnessing real-time monitoring capabilities, phishing detection systems can stay one step ahead of cyber adversaries, enhancing overall cybersecurity resilience in an ever-changing digital landscape.

REFERENCES

1. Karim, Abdul, et al. [1] examine the efficiency of gradient boosting classifiers (GBM) as a machine learning model for identifying phishing URLs.
2. Nowroozi et al. [2] does not explicitly mention the detection of phishing websites using urls with gradient boosters.
3. Abhishek et al. [3] does not specifically mention the use of gradient boosters to detect phishing websites using URLs.
4. Daojing, et al [4] proposes a phishing website detection model based on tiny-Bert stacking, which includes gradient booster (GBDT) as a second-level learner in the classifier.
5. Lichen, et ,al[5] provided paper does not mention using gradient booster for detecting phishing websites using URLs. The paper focuses on a deep learning-based framework for detecting and visualizing online malicious advertisements.
6. Hung et al. [6] 's paper does not mention the use of gradient booster for detecting phishing websites using URLs. The paper is about using deep learning to learn a URL representation for malicious URL detection.
7. Zuquan, et al [7] provided paper focuses on the vulnerability of existing deep learning- based malicious URL detection models and does not explicitly mention the use of gradient booster for

detecting phishing websites using URLs.

8. *Z Peng*, et al [8] provided paper does not explicitly mention the use of gradient booster for detecting phishing websites using URLs. The paper focuses on the vulnerability of existing deep learning-based malicious URL detection models to adversarial samples.
9. Ashraf, et al [10] provided paper does not mention the use of gradient booster for detecting phishing websites using URLs.
10. Ch. Rupa, et al[11]discusses detecting malicious links from the host-based and lexical features of URLs to protect users from identity theft attacks. There is no mention of using gradient booster specifically for detecting phishing websites.
11. Mohammad, et al [12] provided paper does not mention the use of gradient booster for detecting phishing websites using URLs.
12. Dao, et, al [13] provided paper does not mention using gradient booster for detecting phishing websites using URLs.
13. Afiqah,et, al [14]provided paper does not discuss the detection of phishing websites using URLs or gradient booster. The paper focuses on the comparison of machine learning techniques for detecting phishing e-mails.
14. Immadisetti, et, al [15] provided paper does not specifically mention the use of gradient booster for detecting phishing websites using URLs.
15. R.Suriya, et al [16] provided paper does not mention the use of gradient booster for detecting phishing websites using URLs.
16. Suhail, et al [17] provided paper does not mention the use of gradient booster for detecting phishing websites using URLs. The paper focuses on detecting phishing e-mails using selected features and machine-learning approaches.
17. Suhail, et al [18] does not mention the use of gradient booster for detecting phishing websites using URLs. The paper focuses on detecting phishing e-mails using selected features and machine-learning approaches.
18. Ekta, et al[19] provided paper does not mention the use of gradient booster for detecting phishing websites using URLs.
19. Ebubekir, et, al [20] provided paper does not specifically mention the use of gradient booster for detecting phishing websites using URLs.
20. Fatimah, et, al [21] provided paper does not mention the use of gradient booster for detecting phishing websites using URLs.
21. Man, et, al [22] provided paper does not mention using gradient booster for detecting phishing websites using URLs. The paper focuses on using support vector machines, K- nearest neighbors, random forest, and Naive Bayes classifiers for detecting malicious URLs.
22. https://www.kaggle.com/datasets/jayaprakashpondy/phishing-websites-feature-dataset