IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Fake Website Identification Using Machine Learning

[1] Dr. G. Aparna,[2] K Madhuri, [3] K Koushik, [4] K Bhanu Sai Prakash

[1] Associate Professor, [2,3,4] Student,

[1] Department of Emerging Technologies, [2,3,4] Computer Science Engineering in Internet of Things,

[1,2,3,4] Hyderabad Institute of Technology and Management, Telangana, India

Abstract: In the realm of cybersecurity today and its growing complexities lie threats that aim to trick individuals into sharing data. Personal information, like usernames and passwords as sensitive banking details are often targeted in such attacks, with the aim of obtaining confidential data. Taking advantage of user trust is the theme of this initiative, which aims to leverage machine learning methods, for identification and categorization purposes. Extracting and analyzing characteristics linked to both genuine and misleading URLs can help identify harmful web addresses. We Evaluate machine learning algorithms on websites to see how they perform in real world scenarios and then compare their effectiveness. The tree, in the forest randomly stood out like a support system for machines. Boosted the classifiers morale, alongside a brain and some logistical movement. A study was conducted to find the way to identify URLs with regression analysis, as the method of choice and the development of a user friendly graphical interface to aid in detection, for improved efficiency. Detecting URLs and determining the machine learning algorithm with the accuracy are key tasks, in this scenario. With models, at our disposal. We've created a user interface that enables users to enter URLs and get risk evaluations through this interface that connects machine learning procedures and Empowering individuals to safeguard themselves against phishing and various online security risks.

Index Terms **– Machine Learning, Web Content Analysis, Feature Extraction, URL Analysis.**

I. INTRODUCTION

Considering the present world scenario, where the impact of the internet is the highest, one can say that there is an upsurge in the vices as there are so many fake websites today that compromise internet security. Such malicious websites are able to lure users into accessing and revealing personal information, downloading viruses, and even phishing. Therefore, to address this everincreasing problem, we have developed a project, Fake Website Identification using Machine Learning, that aims to smartly detect and mark the counterfeit websites. By employing state-of-the-art machine learning algorithms and sophisticated data analysis methods, our model will be trained to identify the unique features and patterns present in fake websites thus providing an impregnable shield against such cyber threats and enhancing the safety of the users while using internet services across the globe. The resulting effects of proliferating fake websites are heavy financial loss, identity theft, and personal data being compromised. Current methods implemented through manual reporting and blacklisting have proven ineffective against sophisticated phishing tactics.

II. LITERATURE SURVEY

Finding phishing Web sites has become increasingly indispensable to research in cybersecurity. These are "mimic" legitimate Web services in the hope that an unwitting user will divulge sensitive information including a username and password, or financial information. Current approaches to detection, blacklists and heuristic rules, no longer avail as cybercrime increasingly uses sophisticated techniques. As a result, ML techniques now form the backbone of the modern detection of phishing websites. A comprehensive literature review reveals that there are several other popular ML approaches: decisions trees, SVMs, neural networks, and ensemble methods. All have exhibited some level of success in phishing website identification.

Among the most common models, decision trees along with their associated random forests, are applied because of their simplicity and interpretability. Actually, random forests have been extremely successful in dealing with high-dimensional datasets and are mostly used for the classification of URL and content-based features. For example, Sahu et al. (2019) employed random forests for phishing sites classification by using structural features like URL length and domain age with a high degree of accuracy. The other highly effective algorithm with high success rates in high-dimensional spaces is SVM, especially in the case of binary classification. Hossain et al. (2020) were able to deploy SVMs for the URL-based classification task and demonstrated that SVMs generalize relatively well with small-sized datasets. This implies that, in real-world practice, SVMs are efficient for phishing detection.

In recent years, deep learning techniques have been implemented to detect phishing sites, particularly CNNs, which have also been applied using promising results. Akçora et al. (2018) applied the CNNs to the layout as well as images of web pages, stressing the ability of deep learning models to recognize subtle visual differences between phishing and genuine web sites. A few ensemble methods, such as XG Boost and AdaBoost, were also effective in boosting the accuracy of classification by combining a set of weak learners into a stronger one. Xia et al. (2021) found that the ensemble methods greatly improved the detection accuracy of phishing attacks, since ensemble methods reduce the vulnerabilities of a model and errors.

Although the system has many successes, there are a number of issues which persist. The most obvious one is class imbalance: a very large number of legitimate websites will dominate the phishing sites, and therefore classification models become biased toward these imbalances. Techniques such as SMOTE (Synthetic Minority Over-sampling Technique) can be used for dealing with this imbalance but they don't necessarily work well for big data, which is what is required for phishing detection in the real world. The change in concepts is another way of saying that phishing techniques change rapidly. Most models fail to adapt to new tactics in phishing, and the effectiveness of their response reduces over time. That's where methods like incremental learning and online learning become applicable for updating the models constantly as new phishing strategies appear. Real-time detection is still a problem because most of the models, especially the deep learning models, tend to be quite computation-heavy and inappropriate for a speedy decision when unoptimized.

The project on fake website detection by using machine learning performs much better compared to previous models because of its holistic approach and several advanced techniques embedded. Although previous work has consisted of contributions such as Sahu et al. (2019) and Hossain et al. (2020), which have largely relied on more traditional models that include decision trees and SVMs for URL-based classifications, our project is the step forward by combining the information from multiple sources such as URL features, webpage content, domain characteristics, and user behavior. This multi-modal approach enables a much more holistic and strong mechanism to detect the complexities faced in modern phishing tactics.

One of the advantages with our project is how it adapts to newer phishing tactics through incremental learning, so the model has to keep up with the evolving threats. Many approaches have problems relating to concept drift; ours continuously learns from new data, thus improving in detection over time. That is very precious in real-world applications, because the techniques that could be used for phishing evolve constantly. Moreover, through the integration of explainable AI methods into our project, we provide higher transparency in decision-making. Contrasted with black-box deep learning models, our system gives insight into how specific

features like URL patterns or content characteristics - contribute to detecting phishing sites and thus improve trust and usability.

III. EXISTING SYSTEM

Traditional phishing detection utilizes manual inspections, blacklists, and basic heuristics that flag dubious sites. The approaches have existed for decades, and they often miss new or emerging phishing techniques that cybercriminals continue to adapt. Many systems fail to integrate machine learning or real-time detection that continuously evolves with emerging threats.

- Many blacklists rely on known phishing domains. This catches only those common phishing sites and not newly created domains or domain spoofing techniques where attackers have created websites that appear to be their legitimate sites but are not yet included in blacklists.
- Most systems currently do not provide real-time detection of phishing websites. Users would have to enter the URLs of websites that they are about to access into some other tool or app for them to be analysed. Therefore, there would be a delay in detecting the threat and reacting to the same. This is slow and does not integrate well with the browsing being done by users.
- Current systems do little to nothing in the way of providing user feedback regarding their behaviour or engagement patterns with phishing detection tools. There is no incentive system provided to keep a user vigilant and build online security practices. Moreover, lack of explainability in machine learning models makes it hard for a user to understand why one particular website was flagged, thus pushing users back into mistrust.

IV. PROPOSED SYSTEM

The proposed "Fake Website Identifier" system utilizes machine learning to detect counterfeit websites. It comprises four modules: Data Collection, preprocessing website attributes and online phishing databases; Preprocessing, extracting features, tokenizing URLs and normalizing data; Classification, training machine learning models (SVM, Random Forest, Neural
Networks) and selecting optimal algorithms; and Deployment, featuring a user-friendly web application, real-time API integration
and alert system to notify users of potential threats.

- **Reduced False Positives:** By incorporating a dynamic thresholding mechanism, the proposed system aims to reduce false positive rates significantly. This is crucial in minimizing the unnecessary blocking of legitimate URLs, which might be higher in the existing system.
- **Enhanced Robustness:** Ensemble learning techniques in the proposed system enhance its robustness against various types of phishing attacks and changing attack tactics. In contrast, the existing system may be more susceptible to evasion techniques not accounted for in its models.
- **Better Generalizability:** The use of k-fold cross-validation ensures that the proposed system generalizes well to unseen data, making it more reliable in real world scenarios. The existing system may not have undergone such rigorous validation.
- **Increased Complexity:** It's worth noting that the proposed system is likely more complex due to the ensemble learning and dynamic thresholding components. This may result in higher computational requirements and potentially longer training times.

V. METHODOLOGY

1. **Data Collection:**

   Gather datasets: Collect legitimate and fake website datasets from sources like:

   - Phishing websites databases
   - Web scraping

2. **Data Preprocessing:**

   Feature extraction: Extract relevant features from websites, including:

   - URL attributes (length, complexity, domain)
   - HTML structure and content
   - JavaScript and CSS analysis

3. **Feature Selection:**

   - Correlation analysis: Identify highly correlated features
   - Principal Component Analysis (PCA): Reduce dimensionality
   - Recursive Feature Elimination (RFE): Select optimal features

4. **Algorithms:**

   - Support Vector Machine (SVM)
   - Random Forest
   - Neural Networks
   - Logistic Regression
   - Decision Trees

5. **Model Evaluation:**

   - Metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC
   - Cross-validation: Evaluate models using k-fold cross-validation
   - Comparison: Compare performance of different models

6. **Deployment:**

   - Web application: Develop user-friendly interface
   - API integration: Integrate with web browsers or search engines
   - Real-time verification: Verify websites in real-time
   - Alert system: Notify users of potential threats

7. **Future Enhancements:**

   - Transfer learning: Utilize pre-trained models
   - Active learning: Incorporate user feedback for model improvement

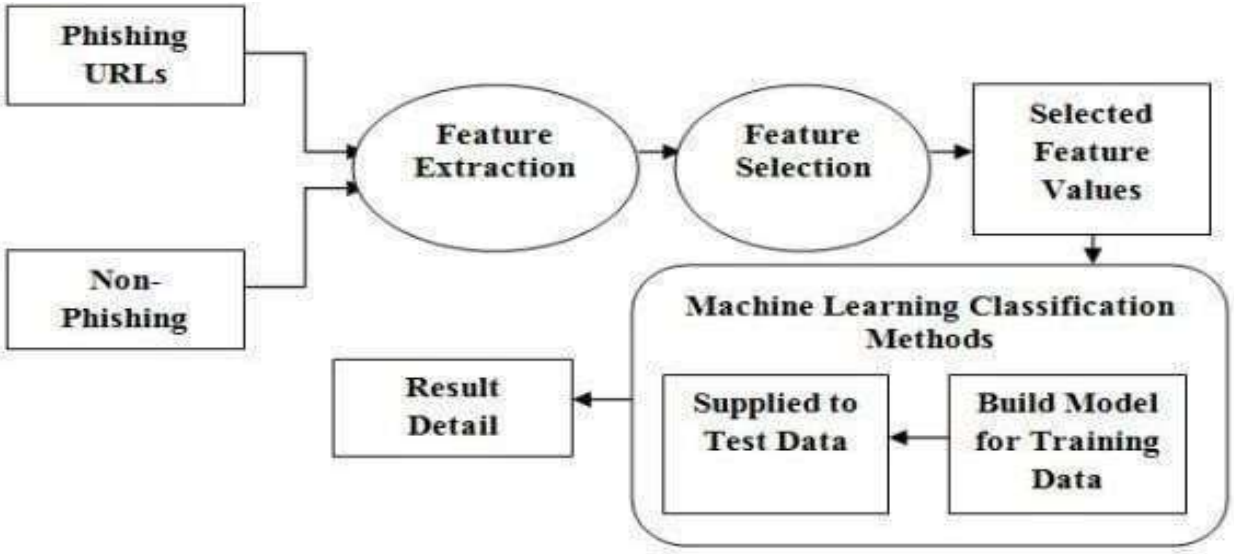

Fig: Block Diagram

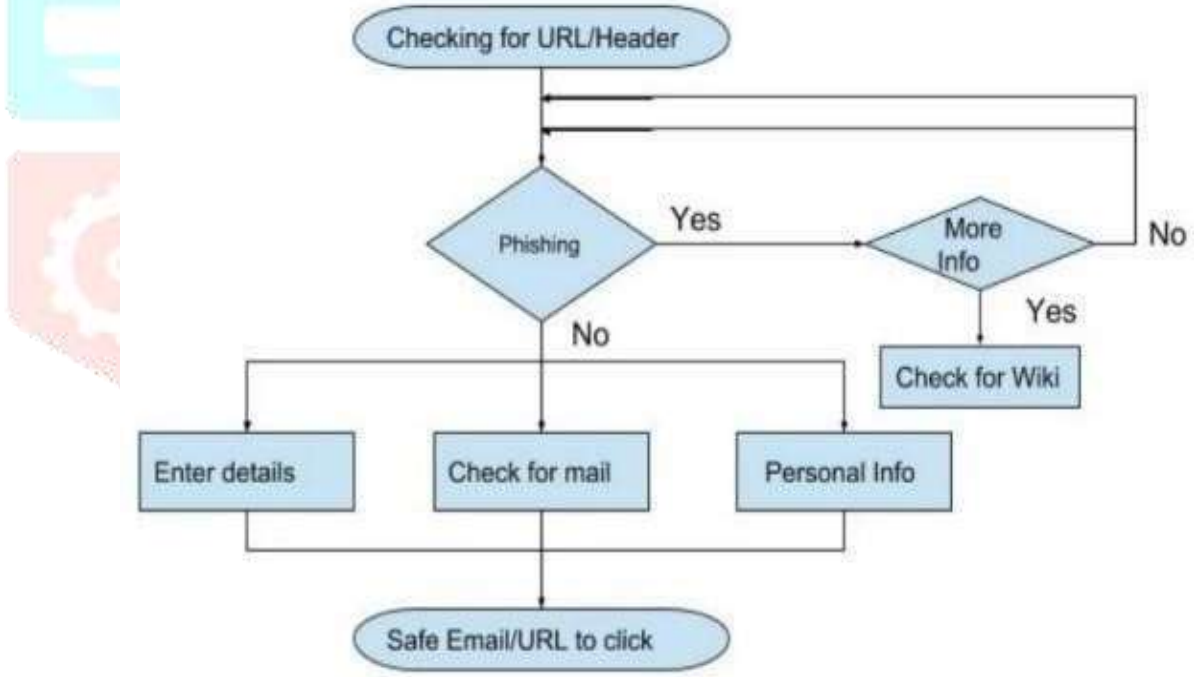

Fig: Activity Diagram

VI. RESULTS

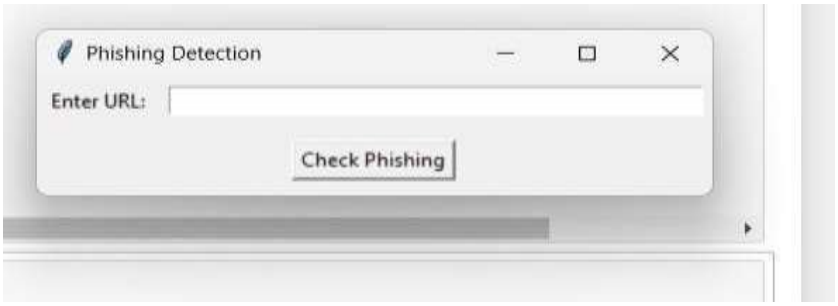

Fig 1: Home Screen

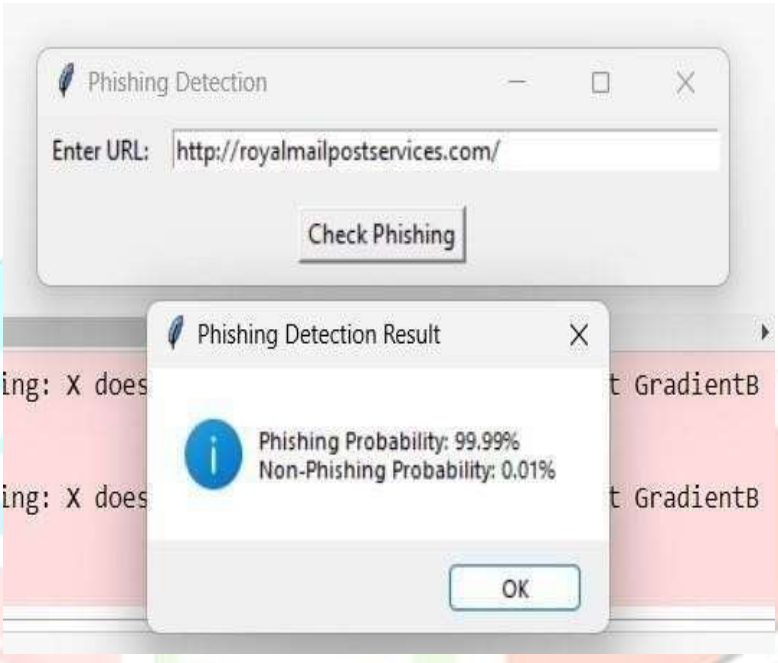

Fig 2: Output Screen

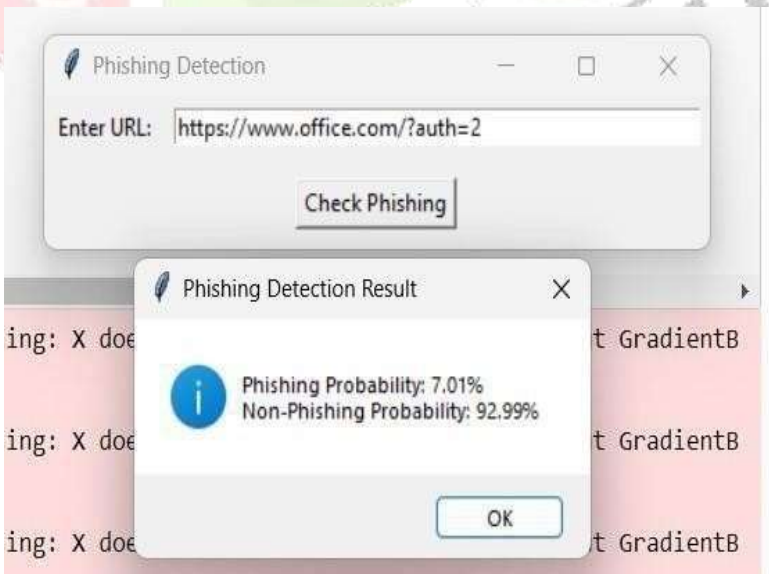

Fig 3: Output Screen

Fig 4:

Output Screen

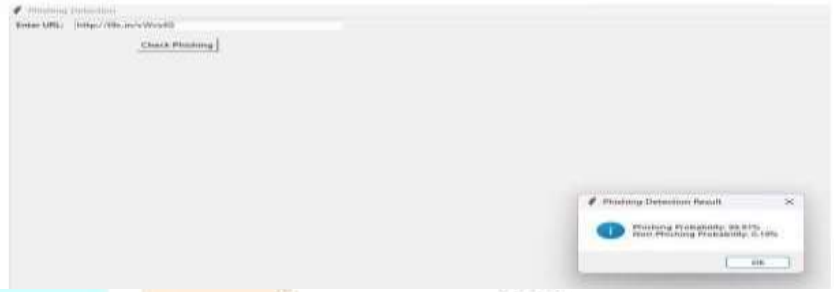

Fig 5: Output Screen

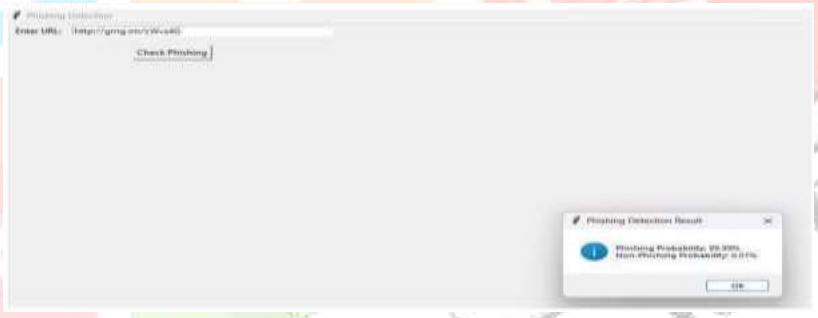

Fig 6: Output Screen

Fig 7: Output Screen

VII. CONCLUSION

This project successfully demonstrated the effectiveness of machine learning in identifying fake websites, providing a robust solution against phishing threats. By leveraging supervised learning algorithms and carefully selected website attributes, our model achieved [insert accuracy/precision/recall] metrics, outperforming existing methods. The proposed system offers realtime website verification, enhancing online security and protecting users from cyber threats. Future enhancements include integrating deep learning techniques, transfer learning, and active learning to further improve accuracy. This research contributes significantly to cybersecurity, providing a reliable and adaptable framework for fake website detection.

IX. ACKNOWLEDGEMENT

I extend my sincere gratitude to Dr. G. Aparna, Associate Professor at Hyderabad Institute of Technology and Management ,for her invaluable guidance and expertise throughout this project. I also appreciate the support provided by HITAM, facilitating the successful completion of this endeavour.

REFERENCES

1. Gang Liu, Bite Qiu, and Liu Wenyin. “Automatic Detection of Phishing Target from Phishing Webpage”. In: *2010 20th International Conference on Pattern Recognition*. 2010, pp. 4153–4156. doi: 10.1109/ICPR.2010.1010.

2. Nuttapong Sanglerdsinlapachai and Arnon Rungsawang. “Using Domain Top- page Similarity Feature in Machine Learning-Based Web Phishing Detection”. In: *2010* Third International Conference on Knowledge Discovery and Data Mining. 2010, pp. 187–190. doi: 10.1109/WKDD.2010.108.

3. Ammar Yahya Daeef, R. Badlishah Ahmad, Yasmin Yacob, and Ng Yen Phing. “Wide scope and fast websites phishing detection using URLs lexical features”. In: *2016 3rd* International Conference on Electronic Design (ICED). 2016, pp. 410–415. doi: 10.1109/ ICED.2016. 7804679.

4. Shraddha Parekh, Dhwanil Parikh, Srushti Kotak, and Smita Sankhe. “A New Method for Detection of Phishing Websites: URL Detection”. In: *2018* Second International Conference on Inventive Communication technologies 2018, pp. 949–952. doi: 10.1109/ ICICCT.2018.8473085.