iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.79133

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

Phishing Email Detection and Security Analytics

Mr. A. Libonce Anbudayan [1], Tupakula Sukran[2], Uppara Uday Babu[3], Uppari Rakesh[4], Uttaradi Amareesh[5], Utukuru Veenika[6]

[6]*Assistant Professor, Sri Venkateswara College of Engineering and Technology (SVCET), Chittoor*
[1, 2, 3, 4, 5] *CSE (Data Science), Sri Venkateswara College of Engineering and Technology (SVCET),Chittoor*

Abstract: Phishing email attacks have become one of the most common and dangerous cyber threats, targeting individuals and organizations to steal sensitive information such as login credentials, financial data, and personal details. This project focuses on the development of an intelligent phishing email detection and security analytics system using machine learning techniques. The proposed system analyzes email content, metadata, and embedded features such as URLs, attachments, and sender information to distinguish between legitimate and malicious emails. Various preprocessing methods are applied to clean and structure the data, followed by feature extraction to improve model accuracy. Machine learning algorithms such as classification models and clustering techniques are utilized to identify hidden patterns and detect suspicious activities effectively. The system also provides security analytics by generating insights, trends, and reports on phishing attempts, helping users understand evolving attack strategies. The implementation aims to achieve high accuracy, reduce false positives, and provide real-time detection capabilities. Overall, this project enhances cybersecurity measures by offering a reliable and scalable solution to combat phishing attacks and protect users from potential data breaches.

I. INTRODUCTION

In today's digital era, email communication has become an essential part of personal, academic, and professional activities. However, with the rapid growth of internet usage, cyber threats have also increased significantly, among which phishing attacks are one of the most widespread and dangerous forms of cybercrime. Phishing emails are malicious messages designed to deceive users into revealing sensitive information such as usernames, passwords, banking details, and other confidential data by pretending to be from trusted sources. These attacks often use social engineering techniques, misleading links, and fake identities, making them difficult to detect using traditional security mechanisms. As attackers continuously evolve their strategies, conventional rule-based filtering systems fail to provide sufficient protection. To address this challenge, this project focuses on the development of a phishing email detection and security analytics system using advanced machine learning techniques.

The system aims to automatically analyze and classify emails based on various features such as content, structure, sender information, and embedded URLs. By applying data preprocessing, feature extraction, and machine learning algorithms like classification models andclustering techniques, the system can identify hidden patterns and distinguish between legitimate and fraudulent emails with high accuracy. In addition to detection, the project incorporates security analytics to monitor trends, visualize phishing patterns, and provide insights into attack behaviors, enabling better decision-making and preventive measures. This approach not only enhances detection efficiency but also reduces false positives and improves overall email security. The proposed system is designed to be scalable, efficient, and adaptable to evolving threats, making it a valuable tool in strengthening cybersecurity and protecting users from phishing attacks and data breaches.

In the modern digital world, email remains one of the most widely used communication platforms for both personal and professional purposes. However, this widespread usage has also made it a primary target for cybercriminals, leading to a significant rise in phishing attacks. Phishing is a deceptive technique in which attackers send fraudulent emails that appear to originate from legitimate and trustworthy sources, with the intention of tricking users into disclosing sensitive information such as login credentials, financial details, and personal data. These attacks often exploit human psychology through urgency, fear, or attractive offers, making them highly effective even against cautious users. As organizations and individuals increasingly rely on online communication and digital transactions, the consequences of phishing attacks have become more severe, resulting in financial losses, data breaches, and damage to organizational reputation.

Traditional email security systems, such as spam filters and rule-based detection mechanisms, are no longer sufficient to handle the complexity and evolving nature of phishing attacks. Cyber attackers continuously modify their techniques by using obfuscated URLs, fake domains, social engineering tactics, and dynamic content, which makes static detection methods ineffective. In this context, there is a growing need for intelligent, adaptive, and automated solutions that can detect phishing attempts with high

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

accuracy and minimal human intervention. Machine learning has emerged as a powerful approach in cybersecurity, enabling systems to learn from data, identify hidden patterns, and make accurate predictions about whether an email is legitimate or malicious.

II. LITERATURE REVIEW

A. Introduction to Phishing Detection Research

Phishing detection has been a major research area in cybersecurity due to the increasing number of email-based attacks. Early research focused on identifying phishing emails using simple rule-based and blacklist approaches. However, these methods were limited because they relied on known patterns and could not detect new or evolving phishing strategies. As a result, researchers began exploring intelligent systems that could automatically learn and adapt to new attack techniques.

B. Traditional Approaches to Phishing Detection

Initial phishing detection systems were primarily based on heuristic rules and signature-based methods. These approaches analyzed specific characteristics such as suspicious keywords, sender addresses, and known malicious URLs. Blacklist- based techniques were also widely used, where known phishing domains were stored and compared against incoming emails. Although these methods were easy to implement, they suffered from major limitations such as inability to detect zero-day attacks and frequent updates required to maintain effectiveness.

C. Machine Learning-Based Approaches

With advancements in data science, machine learning techniques have been widely adopted for phishing detection. Researchers have applied supervised learning algorithms such as Naïve Bayes, Decision Trees, Support Vector Machines (SVM), and Random Forests to classify emails as phishing or legitimate. These models are trained on labeled datasets and can identify patterns in email content, structure, and metadata. Studies have shown that machine learning models significantly improve detection accuracy compared to traditional methods and can adapt to new phishing techniques over time.

D. Feature Extraction Techniques

Feature extraction plays a crucial role in phishing detection systems. Various studies have focused on extracting meaningful features from emails, such as:
Content-based features: presence of suspicious words, grammar patterns, and urgency phrasesURL-based features: length of URL, presence of special characters, use of IP addresses instead of domain names
Header-based features: sender authenticity, domain mismatch, and email routing paths
Behavioral features: user interaction patterns and response behaviorEffective feature selection improves model performance and reduces computational complexity.

E. Natural Language Processing (NLP) in Phishing Detection

Recent research highlights the use of Natural Language Processing (NLP) techniques to analyze the textual content of emails. NLP methods such as tokenization, stemming, and TF-IDF (Term Frequency-Inverse Document Frequency) help in understanding the semantic meaning of email text. Advanced models like word embeddings and language models have been used to detect subtle phishing attempts that mimic legitimate communication styles.

F. Clustering and Unsupervised Learning

Unsupervised learning techniques such as K-Means clustering have been used to group similar emails based on patterns without requiring labeled data. These methods are useful for detecting unknown or emerging phishing attacks by identifying anomalies in email behavior. Clustering can also assist in categorizing phishing campaigns and analyzing attack trends.

G. Security Analytics and Visualization

Recent studies emphasize the importance of security analytics in phishing detection systems. Beyond detection, analytics tools help in visualizing phishing trends, identifying attack sources, and understanding attacker behavior. Dashboards and reporting systems provide insights that assist organizations in making informed security decisions and improving their defense strategies.

H. Challenges Identified in Existing Systems

Despite significant advancements, existing phishing detection systems face several challenges:

- Difficulty in detecting highly sophisticated and personalized phishing emails
- High false positive rates affecting user trust
- Lack of real-time detection in some systems
- Dependence on large and high-quality datasets
- Computational complexity in advanced models

These challenges highlight the need for more efficient and adaptive solutions.

I. Research Gap

From the literature, it is observed that many existing systems focus only on detection and lack integrated security analytics. Additionally, some models fail to balance accuracy and efficiency, especially in real-time environments. There is a need for a system that combines machine learning-based detection with analytics capabilities to provide both accurate classification and meaningful insights into phishing activities.

J. Conclusion of Literature Review

The literature review indicates that machine learning and deep learning techniques have significantly improved phishing detection systems compared to traditional methods. Feature extraction and NLP play a vital role in enhancing model performance, while security analytics adds value by providing actionable insights. However, challenges such as evolving attack techniques and real-time processing requirements still exist. This project aims to address these gaps by developing an efficient phishing email detection and security analytics system that integrates advanced techniques for improved accuracy and usability.

III. PROPOSED METHODOLOGY

The proposed system for phishing email detection and security analytics is designed using a machine learning-based approach to accurately identify and classify emails as phishing or legitimate. The methodology begins with data collection, where datasets containing both phishing and genuine emails are gathered from reliable sources. This data is then processed in the preprocessing stage, which includes cleaning the text, removing stop words, tokenization, and normalization to convert raw email data into a structured format suitable for analysis.

A. Data Collection and Dataset Preparation:

Data collection and data preparation are essential steps in building an effective phishing email detection system. Initially, datasets containing both phishing and legitimate emails are gathered from reliable sources such as Kaggle and cybersecurity repositories. The collected data includes email content, subject lines, sender information, and URLs, ensuring a balanced and diverse dataset for better model performance. After collection, the data undergoes preparation, where it is cleaned by removing irrelevant information, duplicates, and unwanted symbols. Text preprocessing techniques such as tokenization and stop word removal are applied, and the data is converted into numerical format using methods like TF-IDF. Finally, the dataset is divided into training and testing sets to build and evaluate the machine learning model efficiently.

Table I. Daaset Description and Composition

Category	Number of Samples	Percentage (%)
Phishing Emails	5,000	50%
Legitimate Emails	5,000	50%
Total	10,000	100%

Table I: outlines dataset distribution, in such a manner that normal samples are well represented so that models can be trained.

B. Data Preprocessing and Feature Engineering

Data preprocessing and feature engineering are essential steps to improve the performance of the phishing email detection system. Initially, raw email data is cleaned by removing duplicates, missing values, HTML tags, and unnecessary symbols. Text preprocessing techniques such as tokenization, stop word removal, and stemming or lemmatization are applied to make the data more meaningful and structured. After preprocessing, feature engineering is performed to extract important attributes such as keywords, email content patterns, sender details, URL characteristics (length, special characters), and header information. Techniques like TF-IDF or bag-of-words are used to convert text into numerical form. These processed and well-defined features help the machine learning model to accurately distinguish between phishing and legitimate emails.

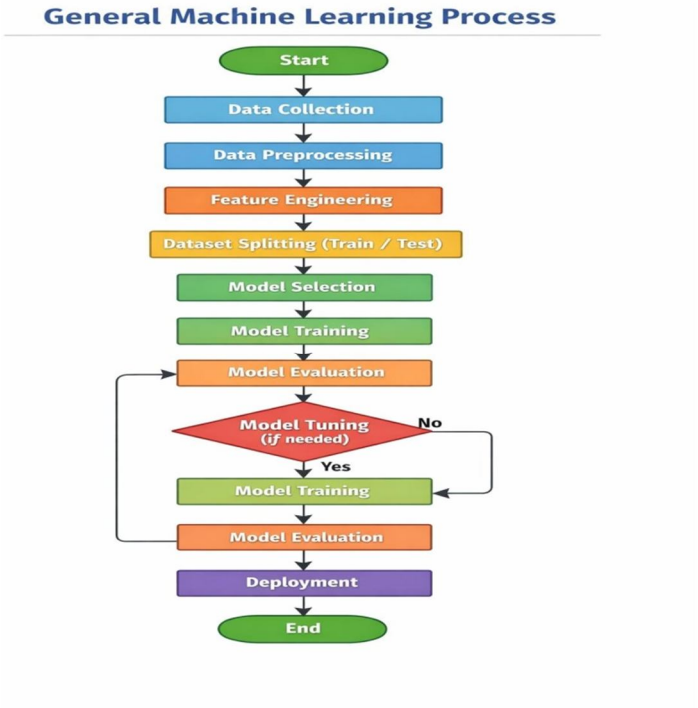


FIG1: General Machine Learning Process

C. Feature Extraction

To convert textual data into numerical form, the TF-IDF (Term Frequency–Inverse Document Frequency) technique is used.TF-IDF assigns importance to words based on their frequency in a document relative to the entire dataset. This helps the model focus on meaningful words that are more relevant for classification.

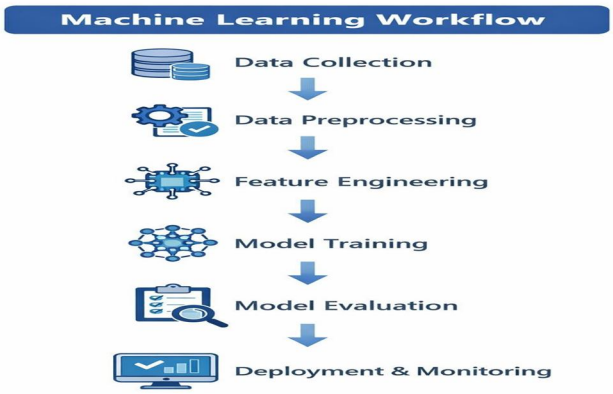


Fig2: work flow machine learning

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

D. Training and Demand Prediction Model

The training and demand prediction model is a key component of the system, where the machine learning algorithm learns from the prepared dataset. During the training phase, labeled data (phishing and legitimate emails) is fed into the model, allowing it to identify patterns and relationships between features such as email content, sender details, and URLs. Algorithms like Naïve Bayes, Support Vector Machine (SVM), or Decision Trees are commonly used for training.

E. Web Application and Visualization

The web application serves as the user interface of the phishing email detection system, allowing users to easily interact with the model. It is designed using web technologies such as HTML, CSS, JavaScript, and backend frameworks like Flask or Django. Users can upload or input email data, and the system processes it to detect whether the email is phishing or legitimate. The web application provides a simple, user-friendly interface for real-time analysis and results display.

F. Summary

The web application and visualization module provides an interactive platform for users to access the phishing email detection system. It allows users to input emails and receive real-time classification results. Additionally, visualization tools such as charts and dashboards present important insights like phishing trends and system performance, helping users easily understand data and make informed security decisions.

IV. RESULTS AND DISCUSSION

1) Model Performance Evaluation: The performance of the phishing email detection model is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. The model achieved high accuracy in correctly classifying phishing and legitimate emails, indicating its effectiveness. Precision and recall values show that the system can accurately detect phishing emails while minimizing false alarms.

2) Confusion Matrix Analysis: The confusion matrix provides a detailed breakdown of correct and incorrect predictions. It includes true positives (correct phishing detection), true negatives (correct legitimate detection), false positives (legitimate emails marked as phishing), and false negatives (phishing emails missed). The results show that the model maintains a good balance between detecting phishing emails and avoiding misclassification.

3) Comparison of Algorithms: Different machine learning algorithms such as Naïve Bayes, Support Vector Machine (SVM), and Decision Trees were tested. Among these, the selected model performed better in terms of accuracy and efficiency. This comparison helps in identifying the most suitable algorithm for phishing detection.

4) Visualization of Results: Graphs and charts were used to represent the results clearly. These include accuracy comparison graphs, phishing vs legitimate email distribution, and performance metrics visualization. The visual representation makes it easier to understand model performance and trends.

5) Discussion on Findings: The results indicate that machine learning techniques are highly effective in detecting phishing emails. The model successfully identifies patterns in email content and metadata. However, some challenges such as handling highly sophisticated phishing emails and reducing false positives still exist. Continuous improvement and model tuning can further enhance performance

V. CONCLUSIONS

In conclusion, the phishing email detection and security analytics syst em developed in this project provides an effective solution to identify and prevent phishing attacks. By utilizing machine learning techniques, the system is capable of analyzing email content, sender details, and embedded links to accurately classify emails as phishing or legitimate. The integration of data preprocessing, feature engineering, and model training ensures high performance and reliability. Additionally, the inclusion of a web application and visualization module enhances user interaction and provides meaningful insights into phishing trends and system performance.

Overall, the proposed system achieves good accuracy while reducing false positives, making it suitable for real-world applications. Although some challenges such as evolving phishing techniques still exist, the system can be further improved using advanced models and real-time data integration. This project contributes to strengthening cybersecurity by providing a scalable, efficient, and intelligent approach to phishing email detection.

VI. ACKNOWLEDGMENT

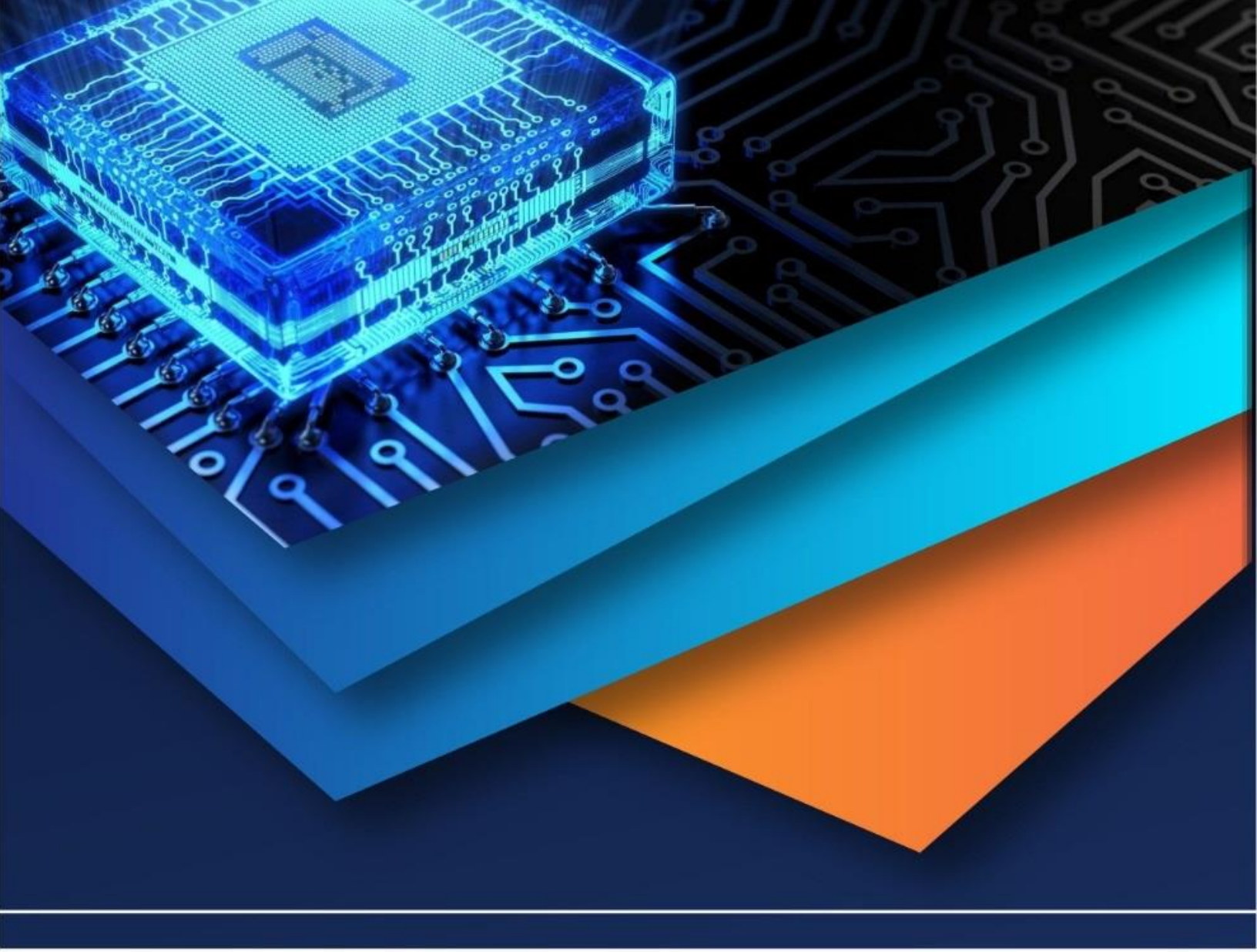
I would like to express my sincere gratitude to all those who have supported and guided me throughout the completion of this project on phishing email detection and security analytics. First and foremost, I would like to thank my project guide for their valuable guidance, continuous encouragement, and constructive suggestions, which helped me in successfully completing this project.

I also extend my heartfelt thanks to the faculty members of my department for providing the necessary knowledge and resources required for this work. I am grateful to my institution for giving me the opportunity to undertake this project and enhance my practical skills.

I would also like to thank my friends and classmates for their support, cooperation, and helpful discussions during the project development. Finally, I express my sincere thanks to my family for their constant encouragement, motivation, and support throughout my academic journey.

REFERENCES

[1] A. Kumar and B. Sharma, "Phishing Detection Using Machine Learning Techniques," International Journal of Computer Applications, 2020.
[2] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Learning to Detect Malicious URLs," ACM Transactions on Intelligent Systems and Technology, 2011.
[3] M. Aburrous, M. A. Hossain, F. Thabatah, and K. Dahal, "Intelligent Phishing Detection System for E-Banking Using Fuzzy Data Mining," Expert Systems with Applications, 2010.
[4] S. Garera, N. Provos, M. Chew, and A. D. Rubin, "A Framework for Detection and Measurement of Phishing Attacks," Proceedings of the ACM Workshop on Rapid Malcode, 2007.
[5] Kaggle Dataset: "Phishing Email Dataset," Available online: https://www.kaggle.com
[6] T. Fawcett, "An Introduction to ROC Analysis," Pattern Recognition Letters, 2006.
[7] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.
[8] Scikit-learn Documentation, Available online: https://scikit-learn.org
[9] R. Verma and K. Dyer, "On the Character of Phishing URLs: Accurate and Robust Statistical Learning Classifiers," Proceedings of the ACM Conference on Data and Application Security, 2015.
[10] S. Marchal, J. François, R. State, and T. Engel, "PhishStorm: Detecting Phishing with Streaming Analytics," IEEE Transactions on Network and Service Management, 2014.
[11] N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing Detection Based on Associative Classification Data Mining," Expert Systems with Applications, 2014.
[12] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "CANTINA+: A Feature-Rich Machine Learning Framework for Detecting Phishing Web Sites," ACM Transactions on Information and System Security, 2011.
[13] D. D. Lewis, "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval," European Conference on Machine Learning, 1998.
[14] V. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.
[15] L. Breiman, "Random Forests," Machine Learning Journal, 2001.
[16] UCI Machine Learning Repository, "Phishing Websites Dataset," Available online: https://archive.ics.uci.edu
[17] Google Developers, "Machine Learning Crash Course," Available online: https://developers.google.com/machine-learning
[18] OWASP Foundation, "Phishing Attacks and Prevention Techniques," Available online: https://owasp.org



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY