ISSN No:-2456-2165

Leveraging Advanced Machine Learning Techniques for Phishing Website Detection

Vaishali Bhargava[1]
Computer Science
Swami Vivekanand Subharti University
Meerut UP, India

Sharvan Kumar Garg[2]
Computer Science
Swami Vivekanand Subharti University
Meerut UP, India

Abstract:- Phishing attacks, which target users through fraudulent websites and emails to steal critical information, continue to pose a significant danger to internet security. Traditional ways of detecting phishing websites frequently fall behind the shifting tactics used by fraudsters. Machine learning approaches are being used as a powerful tool for improving phishing detection capabilities in this kind of context. The current study investigates a novel Machine Learning Model for Detecting Phishing Websites that employ advanced algorithms and feature selection methodologies. The present research analyses the effectiveness of machine learning approaches : J48, and Random Forests under three distinct testing scenarios: a 50% training / 50% testing split, an 80% training / 20% testing split, and a 10-fold cross-validation setup. Through a rigorous experimental approach, the study evaluates their performance using key metrics including accuracy, precision, and recall. Out of the various performance metrics evaluated, J48 demonstrated superior performance with an 80% data split and 10-fold cross-validation. Conversely, when the data was split 50-50, Random Forest yielded better results.

Keywords:- *Phishing; Classification; Decision Tree; Machine Learning; Cyber Security.*

I. INTRODUCTION

Concerns about security have become vital in the rapidly growing digital landscape, during which the internet functions as a core component of several essential operations. Phishing attacks, a fraudulent approach employed by cybercriminals to fool people into disclosing confidential information, pose a serious danger to online security. Phishing websites masquerading as reputable platforms trick naïve users into disclosing sensitive information, resulting in financial losses and compromising privacy.

Due to the rising sophistication of these fake sites, traditional ways of identifying phishing websites frequently fall short. Machine learning, an artificial intelligence branch, has been recognised as a promising approach for detecting and managing such cyber threats. Machine learning models have the ability to analyse trends and traits linked with phishing websites using modern algorithms and data-driven methodologies, allowing for reliable and efficient detection.

This study article goes into the world of cyber security, focusing on the use of algorithms based on machine learning for detecting phishing websites. The goal of this work is to improve the efficiency of phishing detection technologies by exploiting the analytical capabilities of machine learning techniques, thereby bolstering online security measures.

In the following sections, we will look at the current challenges in detecting phishing websites, review relevant literature and strategies, implement the machine learning methodologies mentioned in this paper, show outcomes, and conclude with the significance and future directions of this research. We hope to create a strong and flexible strategy to counter the ever-changing landscape of phishing assaults by harnessing the creative features of machine learning, thereby ensuring an even more secure online environment for people globally.

II. LITERATURE REVIEW

In this study [1], machine learning techniques are employed to identify phishing URLs by extracting and analyzing diverse features from both legitimate and phishing URLs. Decision Tree, Random Forest, and Support Vector Machine algorithms are utilized for phishing website detection. The objective of the paper is to identify phishing URLs and determine the most effective machine learning algorithm by comparing accuracy rates, false positives, and false negatives across each algorithm.

The researchers introduced a multidimensional feature approach [3] for phishing detection, employing a swift detection method through deep learning. Initially, character sequence features from the provided URL are extracted for rapid classification using deep learning, a step that doesn't rely on external help or prior phishing knowledge. Subsequently, these character sequence features are combined with URL statistical features, webpage code features, webpage text features, and the rapid classification output from deep learning, creating multidimensional features for comprehensive analysis.

The researchers developed a Phishing Classification system [4] specifically designed to bypass common phishing detection methods. They utilized numeric representation and conducted a comparative analysis involving classical machine learning techniques such as Random Forest, K nearest neighbors, Decision Tree, Linear SVC classifier, One-class SVM classifier, and wrapper-based feature selection. These

methods involved extracting metadata from URLs to determine the legitimacy of websites.

Research paper [5] presents an up-to-date review of techniques employed in detecting phishing websites. It begins by exploring the phishing life cycle, discusses prevalent anti-phishing methods, with a primary emphasis on identifying phishing links. The focus then shifts to a comprehensive examination of machine learning-based solutions, covering aspects like data collection, feature extraction, modeling, and performance evaluation. The paper extensively compares diverse solutions for detecting phishing websites, providing a detailed analysis of each.

The research [6] provides a deeper insight into various phishing website detection methods, the datasets utilized, and a comparative analysis of algorithmic performance. Machine learning techniques were predominant, with 57 studies employing them, as indicated in the systematic literature review. Data collection primarily sourced from PhishTank (utilized by 53 studies for phishing datasets) and Alexa's website (used by 29 studies for legitimate datasets). Notably, Random Forest Classifier was the most utilized method, chosen in 31 studies. Additionally, Convolutional Neural Network (CNN) emerged as highly effective, achieving an accuracy rate of 99.98% in different studies for detecting phishing websites.

III. PROPOSED MODEL

The website dataset was meticulously pre-processed before undergoing training. After the completion of pre-processing steps, the data was trained using machine learning algorithms, specifically J48 and Random Forest. The outcomes were assessed through three different testing methods: 10-fold cross-validation, as well as evaluations based on data split ratios of 50:50 and 80:20. The workflow is visually represented in Figure 1 (refer to Fig. 1).

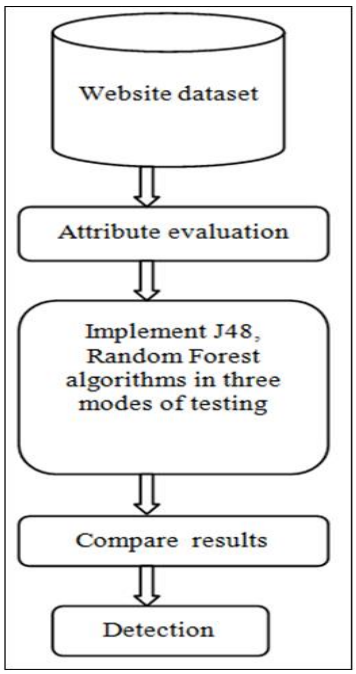


Fig 1 Framework for Experimentation

IV. DATA DESCRIPTION

The dataset used in this experiment is website dataset which is downloaded from UCI machine repository. Dataset contains 1353 instances and, 10 attributes. The URL link for the dataset used is provided in reference [2].

V. DATA PRE-PROCESSING

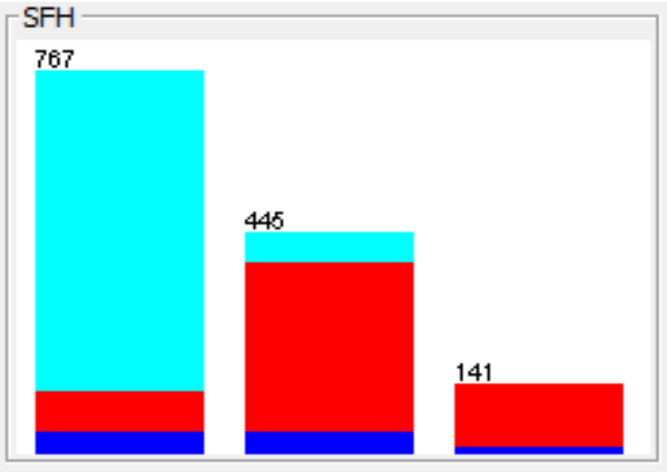
Preparing data is a critical step in predictive modeling, as the quality and structure of input data significantly influence classifier performance. In this section, we elaborate on the data pre-processing steps we implemented to ensure the accuracy and confidentiality of the datasets used in our research.

- *Attribute Evaluation*

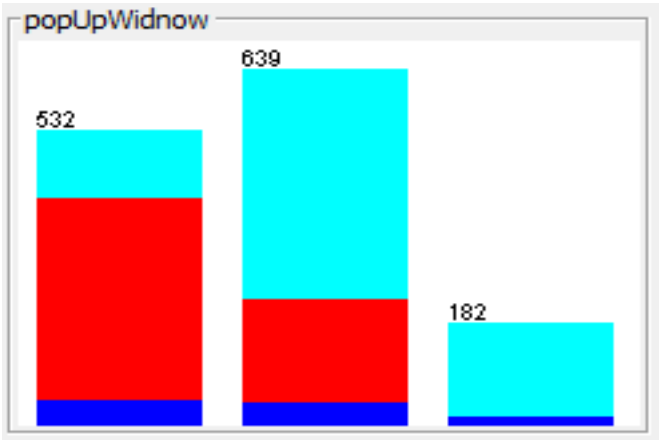
To determine the efficacy of a phishing detection strategy, extensive evaluation metrics are required. It is difficult to select relevant measurements and ensure that the model works adequately across multiple parameters. In this experiment, we employed an information gain-based selection algorithm to assess feature ranks, identifying the most pertinent attributes for constructing performance models. During the feature selection process, each attribute was assigned a rank value, reflecting its influence on data classification. Attributes with the highest ranks were selected, while others were excluded.

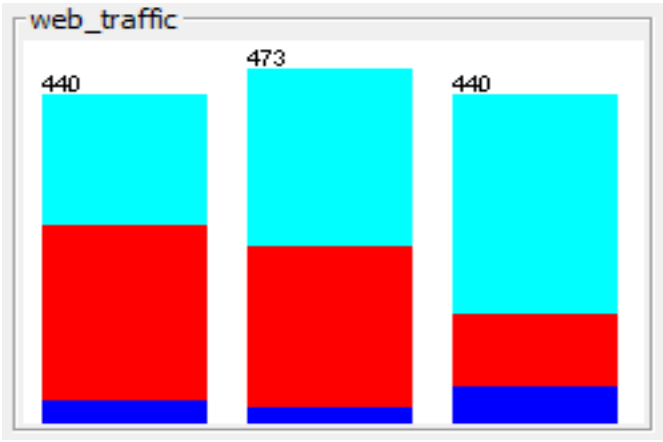
- *Data Vizualization*

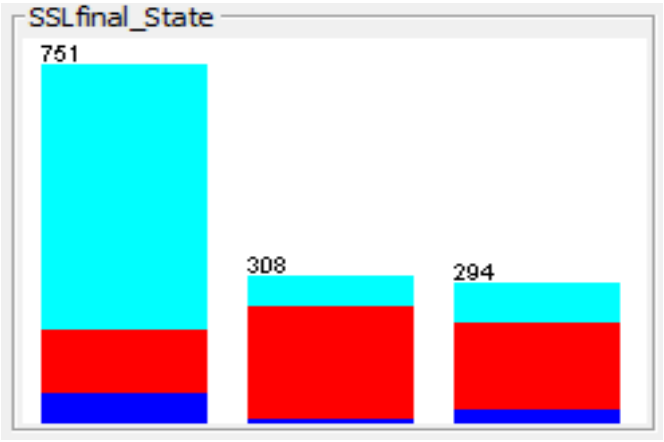
Data visualization is a crucial pre-processing step that employs graphical representation to aid users in understanding and interpreting intricate data. Recently, visualization techniques have been applied to illustrate facets of online learning. Educators can utilize graphical representations to gain a better understanding of their students and monitor activities in remote classes. In this study, the attributes were visualized using the Weka tool. Figure 2 illustrates the visualization of features that received higher ranks during feature evaluation (refer to Fig. 2).

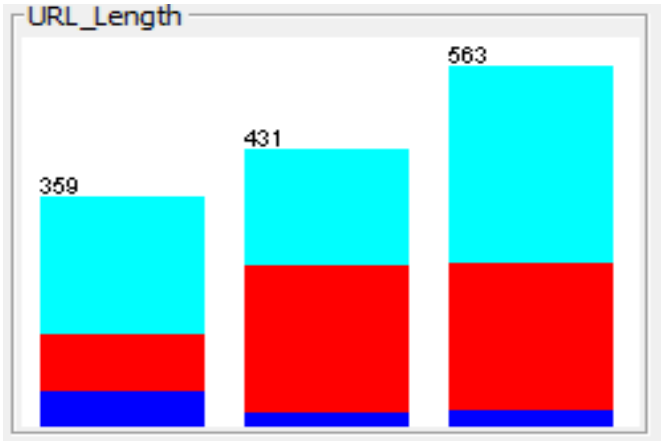


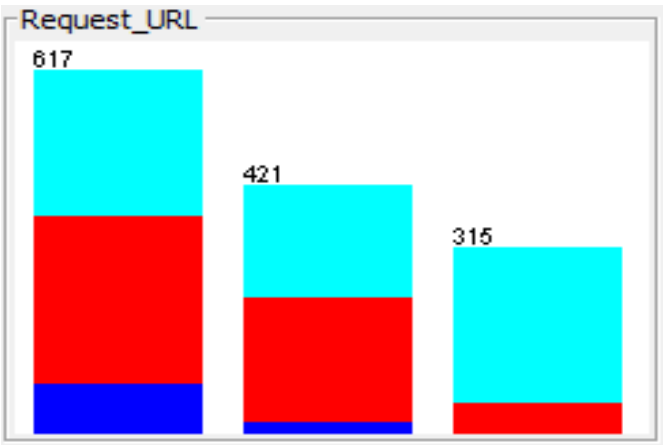
ISSN No:-2456-2165

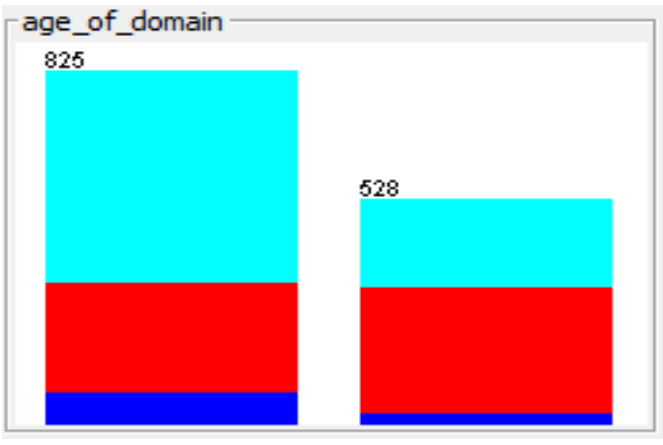


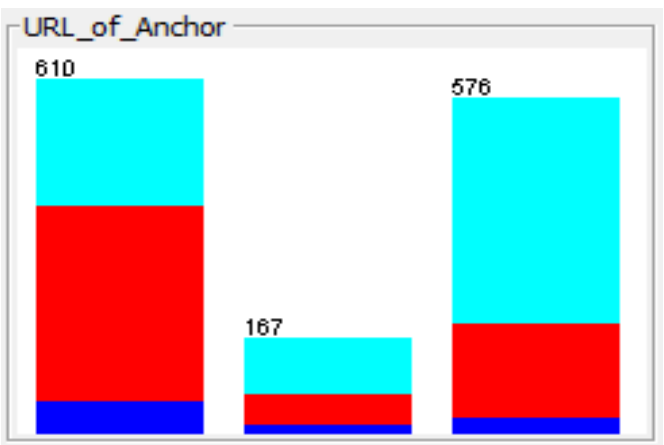


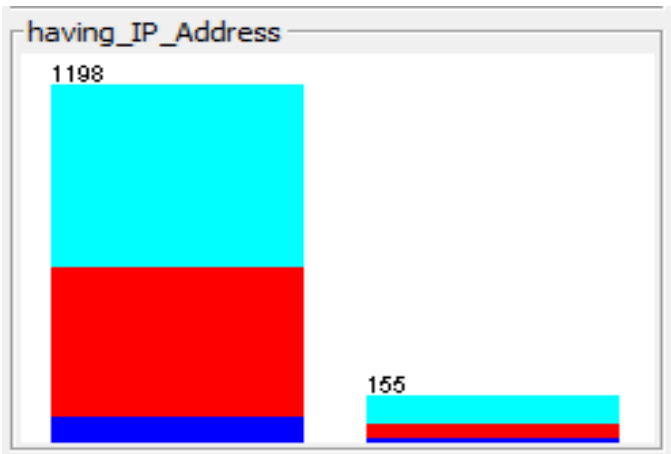












ISSN No:-2456-2165

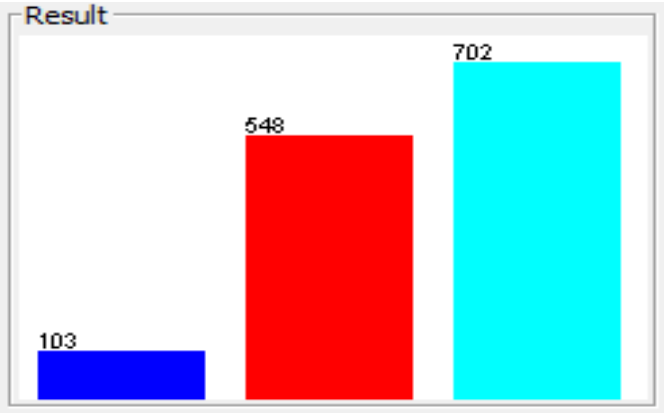


Fig 2 Visualization of the Attributes.

VI. DESCRIPTION OF THE ALGORITHM USED

➢ *J48*

This technique generate a tree by iteratively dividing the dataset depending on the most relevant features, aiming to create an accurate predictive model for class labels. J48 employs a recursive top-down approach, starting at the root node with the entire dataset and then progressively splitting the data based on the attribute that provides the best separation of classes. Renowned for its efficiency, simplicity, and ability to handle noisy data, J48 has found applications in diverse fields, including healthcare diagnosis, finance, and text classification.

➢ *Random Forest*

Random Forest is a powerful ensemble technique in machine learning applicable to tasks such as classification and regression. It extends the decision tree method by incorporating bagging and random feature selection strategies, enhancing prediction accuracy, robustness, and generalization. This method generates a number of decision trees, individually trained on a different subset of the data generated through random sampling with replacement. During the tree construction, a randomly selected set of attributes is considered for each tree, mitigating overfitting and ensuring diversity among the trees. Widely employed in predictive modeling, Random Forest excels in preventing overfitting, managing complex data interactions, and providing insights into feature relevance. It accommodates both qualitative and numerical features, making it versatile for both classification and regression problems.

VII. RESULTS AND DISCUSSIONS

Table 1, 2, 3 shows the compared outcomes in terms of accuracy, precision and recall respectively. Figures 3 and 4 depict the outcomes of the performance metrics analysis. Among the metrics assessed, J48 exhibited superior performance with an 80% data split and 10-fold cross-validation. Conversely, when the data was divided equally (50-50), Random Forest produced better results (refer to Fig.3,4).

Table 1 Performance of Classifiers using 50% Training / 50% Testing

Dataset split	Classifiers	Accuracy	Precision	Recall
50% Training	J48	88.90%	0.888	0.889
	Random Forest	90.23 %	0.903	0.902

Table 2 Performance of Classifiers using 80% Training / 20% Testing

Dataset split	Classifiers	Accuracy	Precision	Recall
80% Training	J48	89.66 %	0.899	0.897
	Random Forest	88.92	0.891	0.889

Table 3 Performance of Classifiers using 10 Fold Cross Validation

Cross Validation	Classifiers	Accuracy	Precision	Recall
10 Fold	J48	90.76 %	0.908	0.908
	Random Forest	89.94%	0.900	0.899

ISSN No:-2456-2165

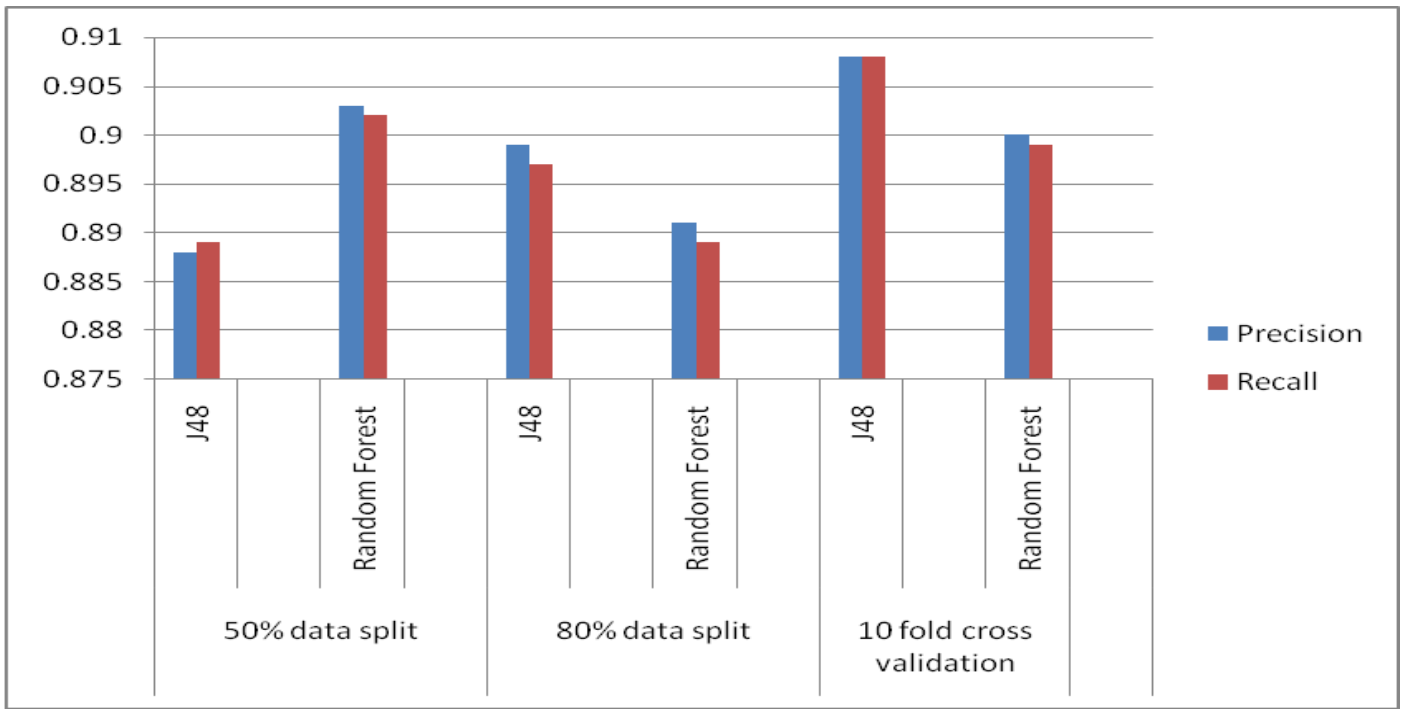


Fig 3 Precision and Recall Comparison of Algorithm in Various Testing Modes.

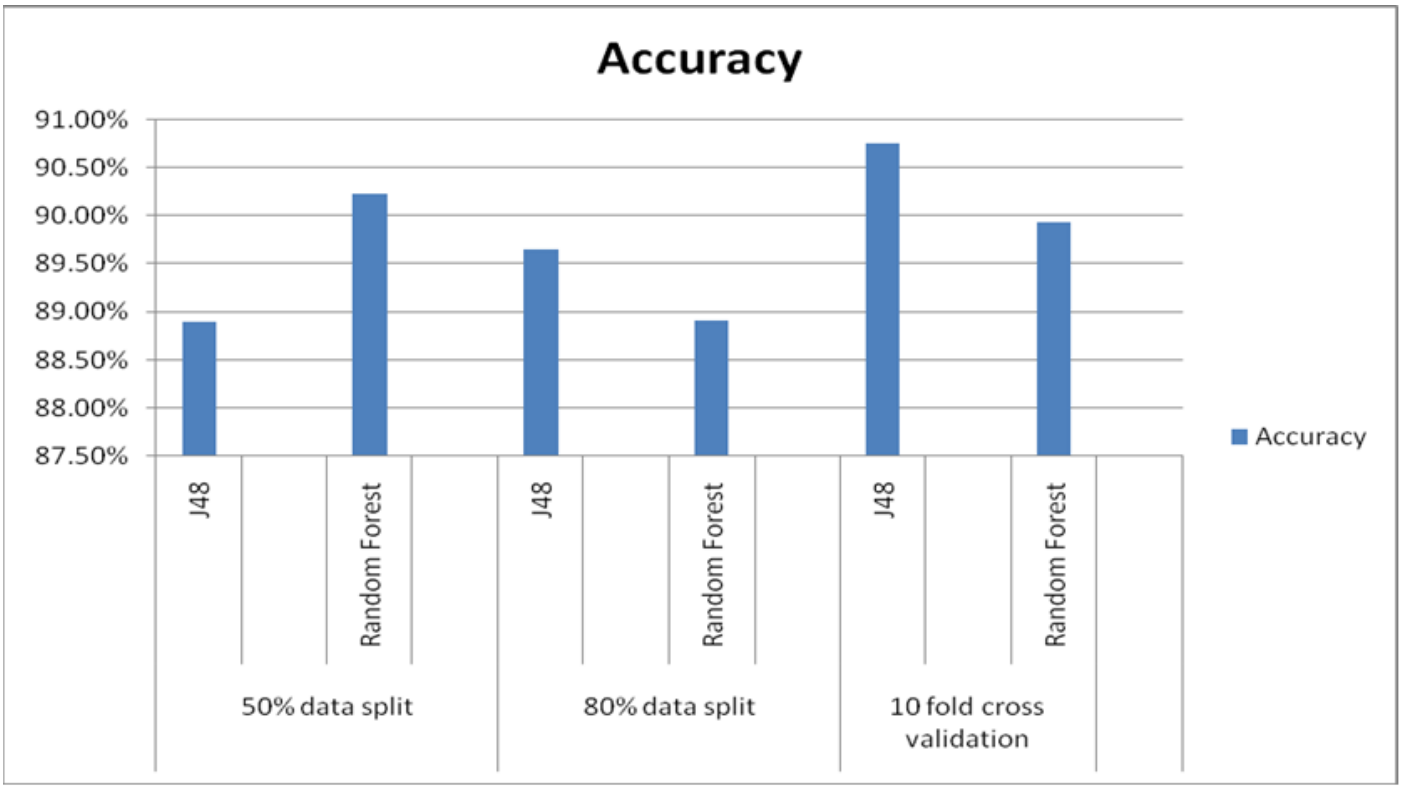


Fig 4 Comparison of Algorithm Accuracy in Various Testing Modes.

VIII. CONCLUSION

In conclusion, this research focused on the recognition of website phishing attacks using machine learning algorithms. The efficiency of the J48 and Random Forest algorithms is evaluated under various testing circumstances using comprehensive experimentation and analysis. J48 demonstrated outstanding accuracy, precision, and recall in the 80% data split and 10-fold cross-validation scenario, demonstrating its supremacy in this unique situation. Random Forest, on the other hand, performed admirably when the data was divided evenly (50-50), presenting a remarkable alternative in cases where data division is more balanced.

Furthermore, our study emphasized the critical role of data preprocessing and feature selection in enhancing classifier performance. Through careful consideration of

feature relevance and meticulous data visualization, we ensured the accuracy and reliability of our results.

While J48 and Random Forest proved effective in our experiments, it is essential to acknowledge the dynamic nature of phishing techniques. Constant evolution in phishing strategies necessitates ongoing research and adaptation of detection algorithms to stay ahead of cyber threats. This study lays the groundwork for future endeavors in the ever-evolving landscape of cyber security, offering valuable insights for researchers and practitioners aiming to bolster online security through machine learning innovations.

REFERENCES

[1]. R. Mahajan, and I. Siddavatam, " Phishing website detection using machine learning algorithms," International Journal of Computer Applications, 181(23), pp.45-47, 2018.
[2]. Abdelhamid,Neda. (2016). Website Phishing. UCI Machine Learning Repository. https://doi.org/ 10.24432/C5B301.
[3]. P. Yang, G. Zhao and P. Zeng, "Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning," in IEEE Access, vol. 7, pp. 15196-15209, doi: 10.1109, 2019.
[4]. G. Harinahalli Lokesh, and G. BoreGowda, "Phishing website detection based on effective machine learning approach," Journal of Cyber Security Technology, 5(1), pp.1-14, 2021.
[5]. L. Tang, and Q.H. Mahmoud, "A survey of machine learning-based solutions for phishing website detection," Machine Learning and Knowledge Extraction, 3(3), pp.672-694, 2021.
[6]. A. Safi, and S. Singh, "A systematic literature review on phishing website detection techniques" Journal of King Saud University-Computer and Information Sciences, Volume 35, Issue 2, 2023, Pages 590-611, ISSN 1319-1578