



(REVIEW ARTICLE)



## Comparative analysis of machine learning algorithms for phishing website detection

Kumaraswamy S<sup>1</sup>, Saishravan Nitish Nayak<sup>2,\*</sup>, Vinodh Kumar N<sup>2</sup> and Mohammad Waseem<sup>2</sup>

<sup>1</sup> Department of Computer Science and Engineering, University Visveswaraya College of Engineering, India.

<sup>2</sup> University Visveswaraya College of Engineering, India.

International Journal of Science and Research Archive, 2024, 12(01), 293–298

Publication history: Received on 25 March 2024; revised on 01 May 2024; accepted on 04 May 2024

Article DOI: <https://doi.org/10.30574/ijrsra.2024.12.1.0796>

### Abstract

As phishing assaults continue to pose a serious hazard in the digital world, trustworthy detection techniques are required. The effectiveness of machine learning techniques in detecting phishing websites is investigated in this study. The best-performing models were XGBoost and Multilayer Perceptrons (MLPs), which obtained test data accuracy of 90.4% and 90.3%, respectively. On the test data, the Random Forest and Decision Tree models showed competitive accuracies of 86.5% and 87.3%, respectively. SVMs, or support vector machines, performed admirably as well, obtaining an accuracy of 86.4% on the test set. Notably, with accuracy of 74.0% on the test data, the Autoencoder Neural Network demonstrated a restricted level of efficacy. These results highlight the effectiveness of XGBoost and MLPs in precisely detecting phishing websites, offering academics and practitioners in cybersecurity useful information.

**Keywords:** Phishing Websites; Machine Learning; Decision Trees; Random Forests; XGBoost; Support Vector Machines; Autoencoder Neural Network; Multilayer Perceptrons

### 1. Introduction

Phishing, one of the most common cyber attacks has become a major problem in recent years abusing consumers' confidence in online contacts with government and financial organisations [1]. Perpetrators, utilise sophisticated means such as SMS, VOIP, faked URLs and counterfeit websites to trick unwary consumers into disclosing important information [1]. These fake websites sometimes resemble authentic ones, making them difficult to detect, and they seek to steal personal information, such as account, passwords, and financial information [1].

Despite extensive attempts to counteract them, phishing assaults continue to be a lucrative, criminal operation, resulting in significant financial losses for organisations and individuals worldwide [2, 3]. Traditional phishing protection method such as blacklisting URLs and heuristic-based detection, have limited success, especially in identifying, zero hour assaults and reducing false positives [3]. As a result, there is an increased interest in using machine learning approaches to improve phishing detection capabilities [3].

This study investigates several approaches of detecting phishing websites, focusing on the use of machine learning algorithms. By examining information collected from both valid and blacklisted URLs, these algorithms hope to increase detection accuracy and reduce the dangers associated with phishing attempts. The following sections discuss the approaches, conclusions, and consequences of research targeted at improving cybersecurity using advanced detection techniques.

\* Corresponding author: Saishravan Nitish Nayak; Email: [saishravan.n.n@campusuvce.in](mailto:saishravan.n.n@campusuvce.in)

### 1.1. Dataset

URLs for benign websites were gathered from [www.alexacom.com](http://www.alexacom.com), whereas phishing URLs were gathered from [www.phishtank.com](http://www.phishtank.com). The data collection contains 50,227 URLs, including 35,378 benign and 14,849 phishing URLs. URLs are labeled as "0" for benign content and "1" for phishing.

---

## 2. Methodology

For the purpose of analysis of different Machine Learning algorithms to detect phishing websites, the first phase was feature extraction. A phishing website is a popular social engineering technique that imitates legitimate uniform resource locators (URLs) and webpages. URLs are marked as (0) for legitimate and (1) for phishing types. The collection of phishing websites was collected from an open source server called PhishTank. Phishing URLs was provided in different formats like csv and json and it got updated hourly. The data was downloaded from the link mentioned below.

“ [https://www.phishtank.com/developer\\_info.php](https://www.phishtank.com/developer_info.php) ”

For legitimate URLs, dataset was downloaded from the link mentioned below.

“ <https://www.unb.ca/cic/datasets/url-2016.html> ” . The number of legitimate URLs in this collection was 35,378.

After downloading the datasets, it was loaded into a DataFrame. The data contained thousands of phishing URLs. But the problem was that, the data got updated hourly.

Without getting into the risk of imbalance, margin value of 10,000 phishing URLs and 10,000 legitimate URLs were considered. 7,500 phishing and 7,500 legitimate URLs were randomly picked from the dataframe. During the feature extraction procedure, several features were extracted from the URLs dataset and classified as address bar-based, domain-based, and HTML/JavaScript-based features. Address bar-based features included domain extraction, IP address checking, '@' symbol presence, URL length, depth, redirection, 'http/https' in the domain name, URL shortening service usage, and domain prefix/suffix '-'. Domain-based functions included examining DNS data, website traffic, domain age, and domain expiration. HTML and JavaScript-based functionality included detecting iframe redirection, customizing the status bar, deactivating right-click, and forwarding the page. These attributes were retrieved using programs designed to examine various properties of URLs, including domain parsing, length computation, and WHOIS database searching. Once retrieved, these characteristics were separated into dataframes for genuine and phishing URLs, which were then concatenated into a single dataframe. This combined dataset of 15,000 URLs was exported for further research, containing 7,500 phishing and 7,500 legal URLs.

The goal of the study was to train machine learning models and deep neural networks to detect phishing websites using a dataset that included both phishing and benign URLs. The procedure began with importing the extracted data and saving it to a CSV file. After loading [7] the data, several approaches were used to comprehend its structure and distribution, such as displaying the data using plots and graphs and assessing relationships with heat maps. Data preparation techniques were then used to clean the data, including removing the 'Domain' column, which had little importance for model training. The data was validated for null and missing values to ensure it was ready for further processing [8].

The data was then separated into features and target columns, denoted as X and y, respectively. The dataset was divided into training and testing sets using an 80/20 split ratio. Several supervised machine learning models were evaluated for training, including Decision Tree, Random Forest, Multilayer Perceptrons (MLPs), XGBoost, Autoencoder Neural Network, and Support Vector Machines (SVMs). Each model was created, trained with the training data, and then assessed for performance using measures like accuracy. In addition, bar plots were used to assess feature relevance for models such as Decision Tree and Random Forest.

For later examination, the performance outcomes of every model on training and testing data were saved. In order to prevent duplication, each set of results was only ever stored once when the results of each model were finally stored. This thorough method of developing and assessing machine learning models gave important new information on how well various algorithms predict phishing websites.

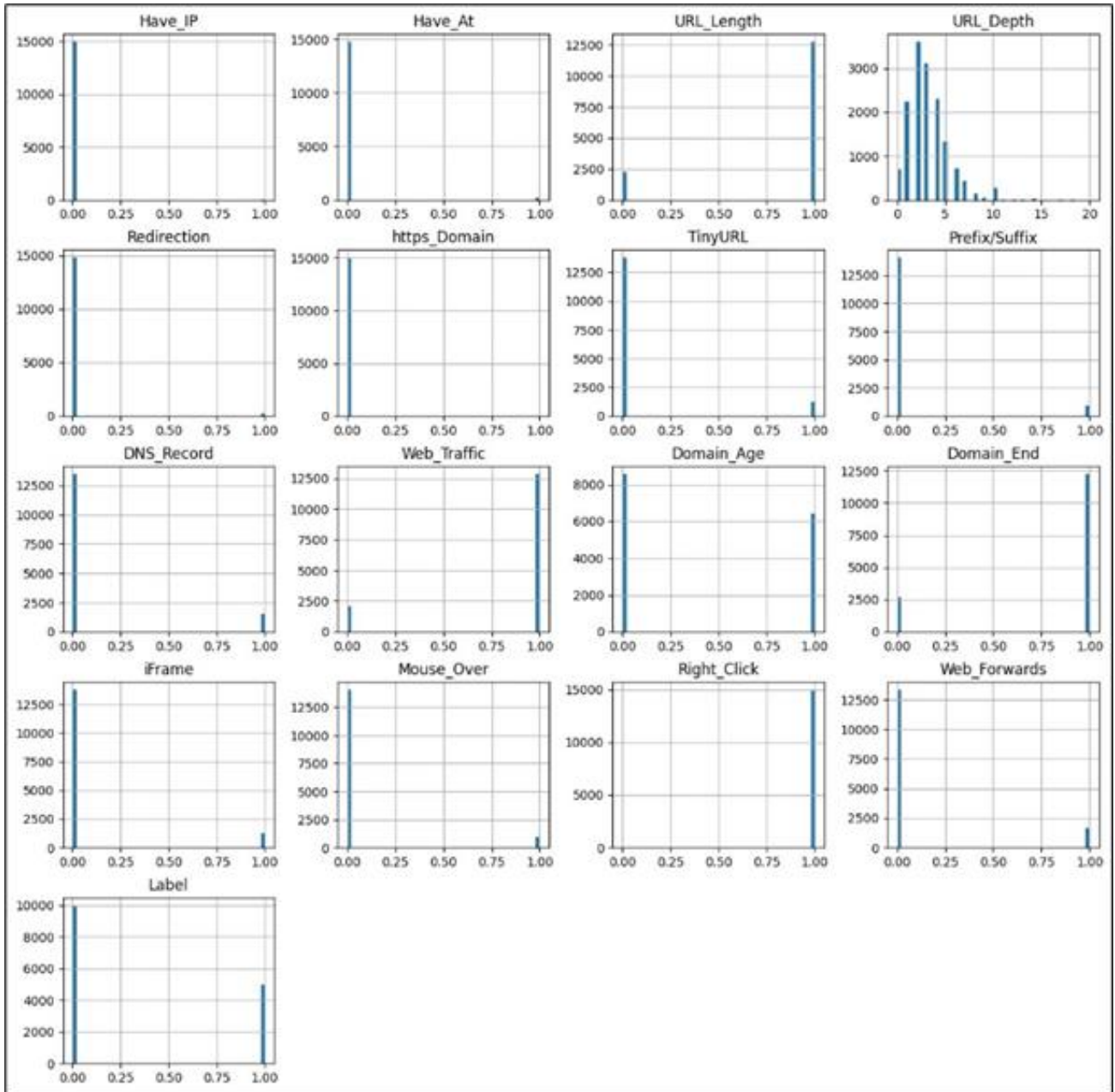
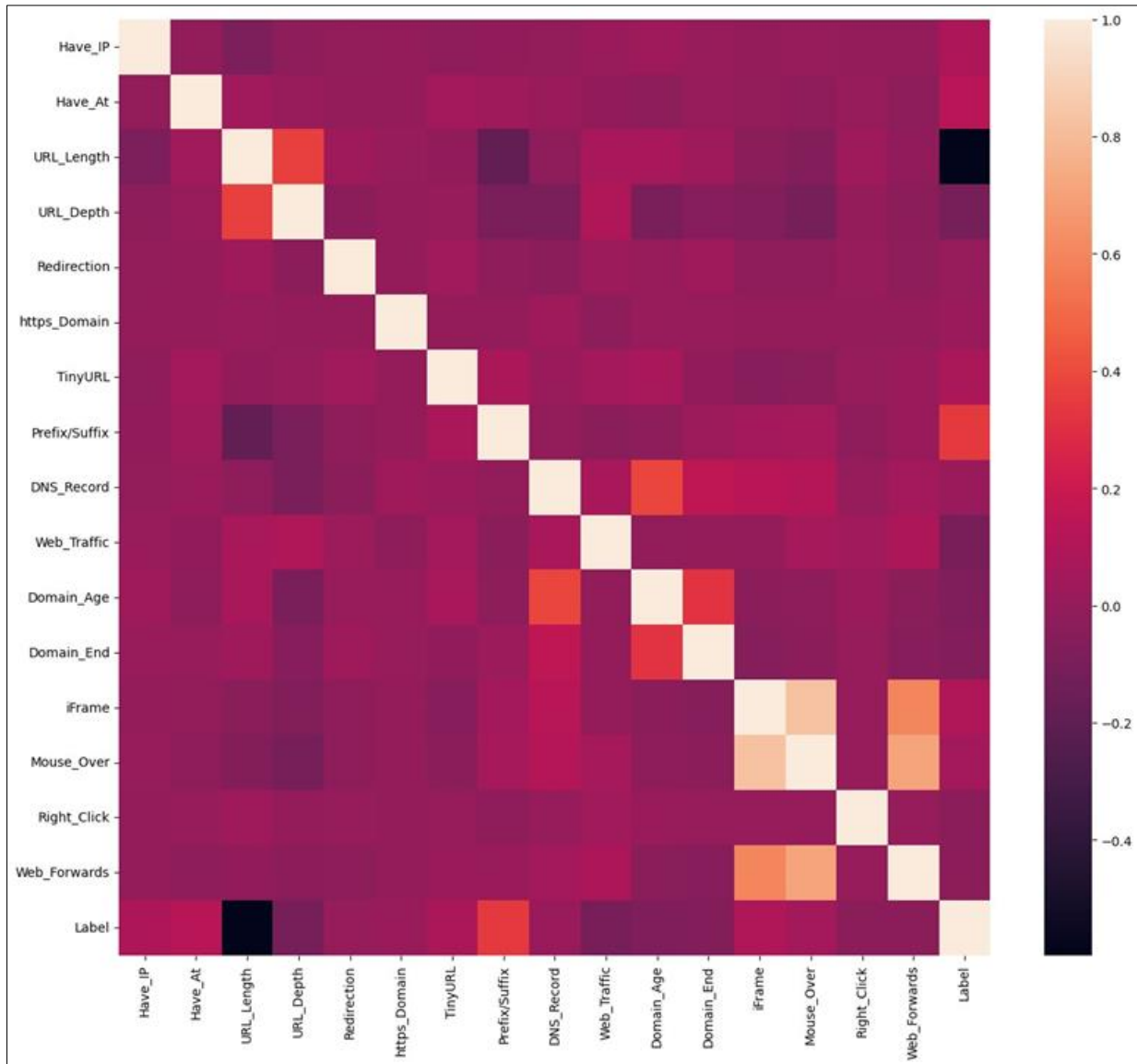


Figure 1 Data Distribution of Dataset among various features



**Figure 2** Correlation Heatmap of Dataset among various features

### 3. Results and Analysis

The project's analysis and findings demonstrate how well different machine learning models anticipate phishing websites. Many classification models were trained and assessed following the dataset's preparation and division into training and testing sets. The effectiveness of different machine learning models for predicting phishing websites based on extracted attributes is assessed in the findings and analysis section. With an astounding accuracy of 90.4% on the test data, XGBoost is the best-performing model. Multilayer Perceptrons (MLPs) come in second with an even more outstanding accuracy of 90.3% on the same dataset. The Random Forest and Decision Tree models perform competitively well, with accuracies of 86.5% and 87.3% on the test data, respectively. Additionally exhibiting strong performance, Support Vector Machines (SVMs) obtain an accuracy of 86.4% on the test data. Though it is implemented, the Autoencoder Neural Network performs noticeably less well, with test data accuracies of barely 74.0%, indicating its limited usefulness. These findings demonstrate how well XGBoost and MLPs detect phishing websites, highlighting their potential value in supporting cybersecurity initiatives.

The results provide researchers and cybersecurity practitioners with useful insights into the subtle differences in performance across various machine learning models when it comes to phishing website identification.

**Table 1** Test Accuracy Result on ML Models with given Dataset

ML Model	Test Accuracy
XGBoost	0.904
Mutlilayer Perceptrons (MLP)	0.903
Decision Tree	0.873
Random Forest	0.865
Support Vector Machine (SVM)	0.864
AutoEncoder	0.740

### Abbreviations

- URLs: Uniform Resource Locators
- CSV: Comma-Separated Values
- EDA: Exploratory Data Analysis
- MLPs: Multilayer Perceptrons
- SVMs: Support Vector Machines

## 4. Conclusion

To sum up, the experiment was successful in assessing how well different machine learning models predicted phishing websites. The outcomes demonstrated the potential of XGBoost and Multilayer Perceptrons (MLPs) for reliable phishing detection systems by demonstrating their efficacy in obtaining high accuracies on both training and test data. These results highlight how crucial it is to deploy cutting-edge machine learning methods to strengthen cybersecurity defences and shield users from internet dangers.

Nonetheless, the Autoencoder Neural Network's comparatively subpar performance indicates that more research into alternate methodologies or optimisation techniques is necessary. This demonstrates the difficulty of detecting phishing emails and the significance of ongoing cybersecurity research and development.

All things considered, the study highlights the importance of continuous innovation in this field and offers insightful information on the capabilities of various machine learning models in thwarting cyber threats. Cybersecurity experts may improve their capacity to identify and counteract phishing assaults by utilising sophisticated algorithms and data-driven strategies. This will protect digital ecosystems and foster confidence and security in online interactions.

## Compliance with ethical standards

### Disclosure of conflict of interest

No conflict of interest to be disclosed.





## References

- [1] Kiruthiga R, Akila D. Phishing websites detection using machine learning. *International Journal of Recent Technology and Engineering*. 2019 Sep;8(2):111-4.
- [2] Kulkarni AD, Brown III LL. Phishing websites detection using machine learning.
- [3] Mahajan R, Siddavatam I. Phishing website detection using machine learning algorithms. *International Journal of Computer Applications*. 2018 Oct;181(23):45-7.
- [4] Machado L, Gadge J. Phishing sites detection based on C4. 5 decision tree algorithm. In 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA) 2017 Aug 17 (pp. 1-5). IEEE.

- [5] Shima K, Miyamoto D, Abe H, Ishihara T, Okada K, Sekiya Y, Asai H, Doi Y. Classification of URL bitstreams using bag of bytes. In 2018 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN) 2018 Feb 19 (pp. 1-5). IEEE.
- [6] Tyagi I, Shad J, Sharma S, Gaur S, Kaur G. A novel machine learning approach to detect phishing websites. In 2018 5th International conference on signal processing and integrated networks (SPIN) 2018 Feb 22 (pp. 425-430). IEEE.
- [7] Abdelhamid N, Ayesh A, Thabtah F. Phishing detection based associative classification data mining. Expert Systems with Applications. 2014 Oct 1;41(13):5948-59.
- [8] Palmer A, Jiménez R, Gervilla E. Data mining: Machine learning and statistical techniques. Knowledge-Oriented Applications in Data Mining, Prof. Kimito Funatsu (Ed.). 2011 Jan 21:373-96.

---

### Authors Short Biography

	<p><b>Kumaraswamy S</b> is currently working as an Assistant Professor in the Department of Computer Science and Engineering, University Visveswaraya College of Engineering, Bengaluru. His research interest is in the area of Data mining, Web mining, Semantic web and cloud computing.</p>
	<p><b>Saishravan Nitish Nayak</b> is currently a fourth-year student at University Visveswaraya College of Engineering, Bengaluru. His interest is in the area of Machine Learning, Cloud Computing, Natural Language Processing, and Big Data.</p>
	<p><b>Vinodh Kumar N</b> is currently a fourth-year student at University Visveswaraya College of Engineering, Bengaluru. His interest is in the area of Internet Of Things, Machine Learning, Quantum Computing, and Artificial Intelligence.</p>
	<p><b>Mohammad Waseem M</b> is currently a fourth-year student at University Visveswaraya College of Engineering, Bengaluru. His interest is in the area of Cyber Security, Machine Learning, and Data Science.</p>