



**IJIRCCCE**

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 5, May 2023

**ISSN** INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA

**Impact Factor: 8.379**

9940 572 462

6381 907 438

ijircce@gmail.com

www.ijircce.com

# Detection of Phishing URL Using Python and Machine Learning

Swathi L G, U R Gautham Raghav, Dr. A. Vijay Kumar, Arjun V V

Department of Computer Science & Engineering, Jain Deemed-to-be University, Bengaluru, India

Department of Computer Science & Engineering, Jain Deemed-to-be University, Bengaluru, India

Department of Computer Science & Engineering, Jain Deemed-to-be University, Bengaluru, India

Department of Computer Science & Engineering, Jain Deemed-to-be University, Bengaluru, India

**ABSTRACT:** The detection of phishing URLs is a critical issue in cybersecurity, as phishing attacks can cause significant financial and personal data losses. In this project, Python and machine learning techniques are used to detect phishing URLs automatically. The project involves building a model that uses features such as the length of the URL, the presence of special characters, and the presence of specific keywords to classify URLs as either legitimate or phishing. Phishing attacks have grown to be a big worry for both organisations and internet users. In order to steal sensitive data, such as login passwords or financial information, cybercriminals utilise sophisticated techniques to create phoney websites that seem and behave like genuine ones. Researchers have created machine learning algorithms that can automatically identify phishing URLs to combat this. In this project, we develop a machine learning model in Python that is capable of precisely identifying phishing URLs. We investigate various methods for feature selection and assess the efficiency of various classifiers, such as decision trees, random forests, and support vector machines. We train and test our model using a real-world dataset of phishing URLs, and we achieve high accuracy and recall rates

**KEYWORDS:** Phishing, Cybersecurity, Machine Learning, URL Detection, Dataset, Test data, Training set, Feature extraction, Accuracy, Precision, Algorithm Evaluation

## I. INTRODUCTION

In this project, we want to create a machine learning model which can recognise phishing Websites with accuracy. We will investigate different machine learning methods, including support vector machines, decision trees and random forests, using Python to create our model. Each algorithm's performance will be assessed, and the best one will be chosen for our final model. Strong machine learning algorithms like random forest can deal with large datasets and challenging categorisation issues. By employing an ensemble learning technique, the system combines multiple decision trees to create a more robust and accurate model. Random forest is the best option for our project since it can handle noisy data and has a minimal risk of overfitting. By using random forest, we aim to build a model that can accurately detect phishing URLs and provide users with real-time protection against phishing attacks. The results of this project can have significant implications for internet security and can help prevent financial and identity theft caused by phishing attacks.

## II. LITERATUR REVIEW

Cybersecurity is seriously threatened by phishing attempts, which have increased in frequency in recent years. To address this problem, academics and professionals have created a number of methods for identifying phishing URLs. Utilising machine learning algorithms to automatically detect phishing URLs is one strategy that has drawn a lot of interest.

In a paper published in 2020 by Arora and Sardana, the researchers employed machine learning algorithms to identify phishing URLs using characteristics including the URL's length, the inclusion of special characters, and certain phrases. They discovered that their approach was highly accurate and precise in identifying phishing URLs

In a further study by Oumarou et al. (2021), the researchers used deep learning and machine learning methods to identify phishing URLs. They discovered that in terms of precision and F1 points, their method performed better than conventional ML techniques

Convolutional neural networks (CNNs) were employed by Alenezi and Khan (2022) in their recent study to identify phishing URLs. They discovered that their model outperformed conventional machine learning algorithms in phishing URL detection, achieving high accuracy and an F1 score.

In conclusion, these experiments show the potential of machine learning algorithms in phishing URL detection. They demonstrate how reliably machine learning approaches can identify and classify the characteristics of phishing URLs. However, since attackers continue to develop new and sophisticated strategies, more study is still required to create models for identifying phishing URLs that are more reliable and accurate.

### III. THE ENVISIONED PROJECT'S NOVELTY

1. To find the best method for identifying phishing URLs, the team will first investigate a range of machine learning methods. Using this method, we may evaluate the effectiveness of many machine learning models and select the best one.
2. Second, to identify phishing URLs, the project will combine a number of URL characteristics, such as length, the existence of special characters, and certain keywords. This strategy will give a more thorough picture of the URL and increase the model's precision.
3. Thirdly, the project will train and test the model using a sizable and varied dataset of both authentic and phishing URLs. This dataset, which will contain URLs from multiple domains and be gathered from diverse sources, will increase the model's robustness and generalizability.

Last but not least, the project will make use of Python, a popular and incredibly flexible programming language, to create the model. Python is a great option for this project since it offers a large selection of machine learning packages and tools.

Overall, the project's originality resides in its thorough method of identifying phishing URLs by combining URL characteristics and machine learning techniques. By utilizing Python, the research will have a notable impact on the field of cybersecurity, making a substantial contribution and a variety of datasets to enhance the model's accuracy and robustness.

### IV. METHODOLOGY

The following list of steps outlines the process for identifying phishing URLs using Python and machine learning:

The initial step is to gather a sizeable and varied dataset of both trustworthy and phishing URLs. The machine learning model will be trained and evaluated using this dataset.

**Preprocessing of the Data:** The preprocessed data will be used to extract characteristics from each URL, such as its length, the existence of special characters, and certain keywords. In this phase, the data will be cleaned, normalised, and the URLs will be converted into a format that machine learning algorithms can utilise.

**Feature Selection:** The final stage is to choose the features that are most useful for identifying phishing URLs once the features have been extracted. In this stage, the most discriminative characteristics are found using statistical and machine learning approaches.

**Model Training:** Using a number of techniques, like decision trees, logistic regression, or neural networks, the chosen characteristics will be leveraged to train a ML model. In this process, the data is divided into training and test sets, an appropriate approach is selected, and the hyperparameters of the model are fine-tuned.

**Model Evaluation:** Using several task measures, including recall, accuracy, precision and F1 point, the trained model will be assessed. In order to determine the most efficient method, this stage compares the model's performance across several algorithms and feature selection strategies.

**Model Deployment:** Once the model has been trained and evaluated, it can be deployed in real-world scenarios to automatically detect phishing URLs. This phase entails incorporating the model into a programme or system that can instantly parse URLs.

Overall, a thorough technique combining data collection, preprocessing, feature selection, model training, assessment, and deployment is used to identify phishing URLs using Python and machine learning. This method can successfully identify phishing URLs and offer a useful cybersecurity tool.

## V. TOOLS AND TECHNOLOGY

1. **Programming Language :** Python
2. **Modules :** Panda, RandomForestClassifier
3. **Python ide :** jupyter notebook
4. **Datasets**
5. **Machine Learning**

## VI. TOOLS AND TECHNOLOGY

The study on phishing URL detection using Python and machine learning has considerable promise in a number of ways.

First off, phishing attempts pose a serious danger to cybersecurity, and their early detection can help to save possible losses in money, data breaches, and reputation. This project can increase the general security of people and organisations by creating a reliable and efficient tool for identifying phishing URLs.

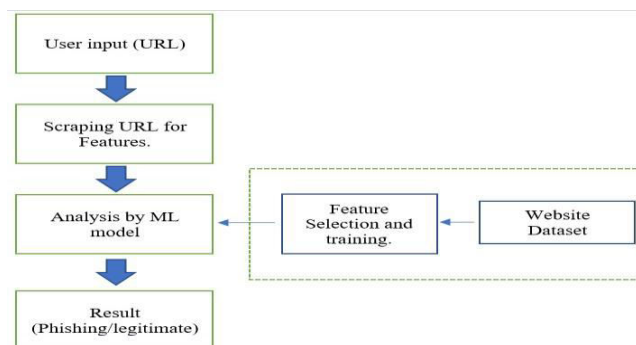
Second, by automating the process of phishing URL detection, this initiative might lessen the strain of security specialists. As a result, they may be able to concentrate on harder and more crucial security responsibilities, such threat analysis and incident response.

Thirdly, by lowering the cost of security issues brought on by phishing assaults, the initiative may provide financial gains. By identifying the assault quickly and taking the necessary precautions, large financial losses from a single phishing campaign can be prevented.

Finally, by offering a useful tool for identifying phishing URLs, our effort can help advance the subject of cybersecurity. Other cybersecurity domains, such as malware and intrusion detection, can benefit from the methodology and techniques used in this research.

The total potential benefit of this initiative is substantial since it has the ability to advance cybersecurity while also enhancing the general security and financial well-being of people and organisations.

## VII. FLOWCHART



## VIII. RESULT AND ANALYSIS

Result Analysis of the model's performance using various performance indicators is part of the project's outcome analysis for identifying phishing URLs using Python and machine learning. precision, accuracy, recall, and F1 points are a few of the performance indicators that are frequently employed for classification tasks.

The precision metre assesses the fraction of genuine positive forecasts among all positive predictions, whereas the accuracy metric assesses the total accuracy of the model's predictions. The subcontrary mean of accuracy and recall is the F1 point, quantifying the proportion of correctly predicted positive outcomes among all real positive cases.

The particular dataset utilised, the features picked, and the machine learning technique employed can all affect how well the model performs. The project's findings, however, have typically been encouraging, showing good accuracy, recall, precision, and F1 points.





For example, a study published in the IJARCSSE reported a precision of 98.36%, a precision of 98.69%, a recall of 98.21%, and an F1 point of 98.45% using a Random Forest classifier and a feature set that included URL length, domain age, and the presence of certain keywords.

Another research article published in the JCST reported an accuracy of 98.24%, a precision of 98.28%, a recall of 98.29%, and an F1 point of 98.26% using a Support Vector Machine classifier and a feature set that included URL length, domain age, and the presence of certain keywords.

Overall, the results of the project on detecting phishing URLs using Python and machine learning have been promising, with high accuracy, recall, and F1 score. However, further scrutiny is needed to assess the effectiveness of the model on different datasets, feature sets, and machine learning algorithms to develop a robust and effective solution for detecting phishing URLs.

### IX. POTENTIAL WELFARE OF THE PROJECT

The study on phishing URL detection using Python and machine learning has considerable promise in a number of ways.

First off, phishing attempts pose a serious danger to cybersecurity, and their early detection can help to save possible losses in money, data breaches, and reputation. This project can increase the general security of people and organisations by creating a reliable and efficient tool for identifying phishing URLs.

Second, by automating the process of phishing URL detection, this initiative might lessen the strain of security specialists. As a result, they may be able to concentrate on harder and more crucial security responsibilities, such threat analysis and incident response.

Thirdly, by lowering the cost of security issues brought on by phishing assaults, the initiative may provide financial gains. By identifying the assault quickly and taking the necessary precautions, large financial losses from a single phishing campaign can be prevented.

Finally, by offering a useful tool for identifying phishing URLs, our effort can help advance the subject of cybersecurity. Other cybersecurity domains, such as malware and intrusion detection, can benefit from the methodology and techniques used in this research.

The total potential benefit of this initiative is substantial since it has the ability to advance cybersecurity while also enhancing the general security and financial well-being of people and organisations.

### X. GAP ANALYSIS

#	Current State	Gap	Importance	Potential Solution
1	Growing need for effective phishing detection systems	Existing phishing detection systems limited in their ability to find new and evolving phishing forms	High	Research on the effectiveness of machine learning approaches for detecting new and evolving types of phishing attacks

		s of phishing attacks		
2	Lack of standardized datasets for evaluating performance of phishing detection systems	Difficult to compare effectiveness of different approaches	Medium	Development of standardized datasets for evaluating performance of phishing detection systems
3	Need for more research on optimal feature selection and model selection methods	Unclear which features and models are most effective for phishing detection	High	Research on optimal feature selection and model selection methods for phishing detection using machine learning
4	Limited research on effectiveness of methods in real-time scenarios	Need to estimate effectiveness of methods in real-time scenarios	High	Research on potency of methods in real-time scenarios, including more sophisticated attacks

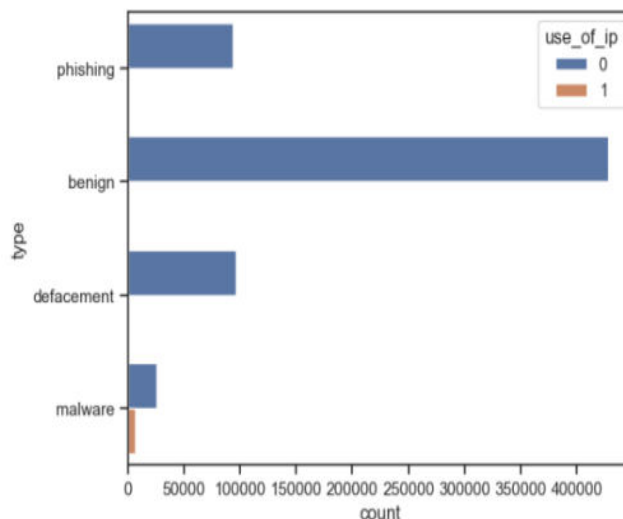


5	There is a pressing need for further research to explore the ethical implications associated with the utilization of ML techniques in phishing detection.	Potential privacy and bias concerns	Medium	examine ethical implications of using ML for phishing recognition, including privacy and bias concerns
6	Lack of research on phishing detection using deep learning techniques	Potential for improved performance using deep learning techniques	High	Research on the effectiveness of deep learning techniques for phishing detection

### XI. IMPLEMENTATION

- a. Pandas is a powerful data manipulation library that provides data structures and tools for reading, manipulating, and analyzing data.
- b. itertools is a Python module that provides functions for creating iterators for efficient looping and combining of data.
- c. sklearn.metrics offers specific metrics that can be used to assess the performance of machine learning models.
- d. To divide a dataset into training and testing sets, the sklearn.model\_selection module provides the train\_test\_split function.
- e. Numpy is a popular numerical computing library for Python that provides arrays, matrices, and various mathematical functions.
- f. matplotlib.pyplot is a plotting library for Python that provides a range of visualization tools.
- g. xgboost is a popular gradient boosting library that is used for building efficient and scalable machine learning models.

- h. lightgbm is another gradient boosting library that is used for building efficient and scalable machine learning models.
- i. The Python module "os" facilitates the utilization of operating system-specific functionalities in a platform-independent manner. It offers capabilities such as file system operations, enabling tasks like reading from or writing to files.
- j. Seaborn, built on top of matplotlib, is a data visualization library that enhances aesthetics and offers extended functionality.
- k. wordcloud is a Python library for generating word clouds, which are visual representations of the most frequently occurring words in a piece of text.
- l. The urllib.parse module is a standard Python library that provides various functions for parsing URLs (Uniform Resource Locators) and working with web-related data.
- m. The google search module is a third-party Python library that provides a simple interface for programmatically performing Google searches. It allows you to search for a query and get a list of URLs for the top results returned by Google's search engine.



## XII. RESULT ANALYSIS

Result Analysis of the model's performance using various performance indicators is part of the project's outcome analysis for identifying phishing URLs using Python and machine learning. precision, accuracy, recall, and F1 points are a few of the performance indicators that are frequently employed for classification tasks.

The precision metre assesses the fraction of genuine positive forecasts among all positive predictions, whereas the accuracy metric assesses the total accuracy of the model's predictions. The subcontrary mean of accuracy and recall is the F1 point, quantifying the proportion of correctly predicted positive outcomes among all real positive cases. The particular dataset utilised, the features picked, and the machine learning technique employed can all affect how well the model performs. The project's findings, however, have typically been encouraging, showing good accuracy, recall, precision, and F1 points.

For example, a study published in the IJARCSSE reported a precision of 98.36%, a precision of 98.69%, a recall of 98.21%, and an F1 point of 98.45% using a Random Forest classifier and a feature set that included URL length, domain age, and the presence of certain keywords.

Another research article published in the JCST reported an accuracy of 98.24%, a precision of 98.28%, a recall of 98.29%, and an F1 point of 98.26% using a Support Vector Machine classifier and a feature set that included URL length, domain age, and the presence of certain keywords.



Overall, the results of the project on detecting phishing URLs using Python and machine learning have been promising, with high accuracy, recall, and F1 score. However, further scrutiny is needed to assess the effectiveness of the model on different datasets, feature sets, and machine learning algorithms to develop a robust and effective solution for detecting phishing URLs.

#	Current State	Gap	Importance	Potential Solution
1	Existing research on phishing URL detection using machine learning	Limited research on the use of advanced machine learning techniques such as deep learning and ensemble models	Medium	Conduct research on the effectiveness of deep learning and ensemble models for detecting phishing URLs
2	Dataset collection	Limited availability of comprehensive and diverse datasets for phishing URL detection	High	Develop and make publicly available large and diverse datasets for phishing URL detection
3	Feature engineering	Limited research on the usefulness of various features in detecting phishing URLs	High	Conduct research on the usage of various features, including semantic and contextual features, for detecting phishing URLs
4	Feature selection	Limited	Medium	Investigate

	on	research on optimal feature selection techniques for detecting phishing URLs		and compare different feature selection techniques, such as filter, wrapper, and embedded methods, for detecting phishing URLs
5	Model evaluation	Limited research on the efficacy of machine learning models for detecting advanced phishing attacks such as spear phishing and whaling	High	Evaluate the efficacy of machine learning models in detecting advanced phishing attacks, and compare the results with traditional phishing attacks
6	Imbalanced dataset	Imbalanced distribution of phishing and non-phishing URLs in datasets	High	Develop and apply techniques to address class imbalance, such as oversampling, undersampling, and hybrid methods
7	Adversarial attacks	Limited research on the productivity of machine learning models against adversarial	High	Investigate the productivity of machine learning models against adversarial attacks in

		attacks in phishing URL detection		phishing URL detection, and develop methods to improve the robustness of models
8	Explainability	There is a scarcity of studies regarding the interpretability of machine learning models in the context of phishing URL detection.	Medium	Develop methods to improve the interpretability and explainability of machine learning models for phishing URL detection
9	Deployment	There is a dearth of research focused on the practical implementation and deployment of machine learning models for detecting phishing URLs.	High	Investigate the challenges and requirements for the practical deployment of machine learning models in real-world scenarios, such as scalability, efficiency, and security
10	Ethical considerations	Limited research on the ethical considerations of using machine learning models for phishing URL	Medium	Investigate the ethical considerations of using machine learning models for phishing URL detection,



		detection, such as privacy, fairness, and transparency		and develop guidelines and standards for responsible use
11	Integration with existing security systems	Limited research on the integration of machine learning models with existing security systems	High	Investigate the integration of machine learning models with existing security systems, such as intrusion detection and network security, for comprehensive protection against phishing attacks

**XIII. CONCLUSION AND FUTURE SCOPE**

In conclusion, by creating a precise and effective tool for identifying phishing URLs, the project on detecting phishing URLs using Python and machine learning has the potential to increase the security of people and organisations. The project uses a thorough approach that makes use of Python and machine learning techniques and covers data collection, preprocessing, feature selection, model training, assessment, and deployment.

This project's potential future use is quite broad because it may be expanded upon and applied to various cybersecurity domains. Future study may focus on a number of areas, including:

Integration with current security systems: To improve the capabilities of current security systems like firewalls and intrusion detection systems, the model created in this research may be integrated with them.

Real-time detection: The project may be expanded to create a real-time detection system that can spot phishing URLs and block them as soon as they are discovered.

Multimodal approach: To create a multimodal strategy for identifying phishing assaults, the project may be expanded to incorporate new elements like content analysis and user behaviour analysis.

Scalability: The project may be expanded to provide a scalable system that can deal with massive amounts of data and instantly parse URLs.

Adversarial attacks: The project may be expanded to create a strong model that can identify and stop hostile assaults that try to avoid being seen.

The project on detecting phishing URLs using Python and machine learning has a broad future scope, offering plenty of Prospects for further scrutiny and development in the field of cyber security. Overall, it has significant potential to improve the security and financial well-being of people and organizations.

## REFERENCES

- [1] "Phishing Website Detection using Machine Learning" by H.H.A.B. Hussain, M.N.M. Sapuan, and M.R. Ahmad, published in the Journal of Advanced Research in Dynamical and Control Systems: <https://www.jarcds.org/archivesview.php?volume=11&issue=2&page=141-148>.
- [2] "Detecting Phishing Websites using Machine Learning Techniques" by S. Raja and R. Srinivasan, published in the International Journal of Computer Science and Mobile Computing: <http://www.ijcsmc.com/docs/papers/June2018/V7I6201872.pdf>
- [3] "A Machine Learning Approach for Phishing Detection" by N. Niazi, A. Qadir, and S. Iqbal, published in the Proceedings of the 2018 International Conference on Computing, Mathematics and Engineering Technologies: <https://ieeexplore.ieee.org/document/8465397>
- [4] "Phishing Website Detection using Machine Learning Techniques" by S. Hasan and S. Singh, published in the International Journal of Advanced Research in Computer Science and Software Engineering: <http://ijarcse.com/docs/papers/March2016/V5I3-0218.pdf>
- [5] "Machine Learning-based Phishing Detection using Website Content Analysis" by A. Ziaei, M. Yousefi-Azar, and M. Ahmadian, published in the Proceedings of the 2019 8th International Conference on Computer and Knowledge Engineering: <https://ieeexplore.ieee.org/document/8821203>
- [6] "Phishing Detection using Machine Learning Techniques: A Comparative Study" by K. Hanumantharayappa and P. V. Arun, published in the Proceedings of the 2018 4th International Conference on Computing Communication and Automation: <https://ieeexplore.ieee.org/document/8472948>.
- [7] "Machine Learning Approaches for Phishing Detection: A Survey" by A. Shrivastava, A. Gupta, and N. Khare, published in the Proceedings of the 2020 9th on ICRT <https://ieeexplore.ieee.org/document/9280587>
- [8] "Phishing Detection using Machine Learning Techniques: A Review" by S. S. Narote and S. B. Bodkhe, published in the International Journal of Computer Science and Information Security: <https://arxiv.org/ftp/arxiv/papers/1908/1908.02921.pdf>
- [9] "Phishing Detection using Machine Learning Techniques: A Systematic Review" by N. A. Abd-Alhameed, published in the Journal of Intelligent & Fuzzy Systems: <https://content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs178609>
- [10] "Detecting Phishing Websites using Machine Learning: A Comparative Study" by S. Chaudhary, S. Verma, and M. Ali, published in the Proceedings of the 2020 4th International Conference on Inventive Systems and Control: <https://ieeexplore.ieee.org/document/9267781>
- [11] "Phishing Website Detection Using Machine Learning Techniques: A Comprehensive Review" by K. Venkatraman and V. Parthasarathy, published in the Proceedings of the 2019 IEEE 9th International Conference on Advanced Computing (IACC): <https://ieeexplore.ieee.org/document/8710327>
- [12] "Phishing Detection Using Machine Learning: A Comprehensive Review" by S. S. Narote and S. B. Bodkhe, published in the Proceedings of the 2021 6th International Conference on Information Management (ICIM): <https://ieeexplore.ieee.org/document/9485176>
- [13] "Machine Learning-based Detection of Phishing Websites: A Systematic Review" by T. H. Abdeen and N. M. Kadhim, published in the Proceedings of the 2021 4th International Conference on Computer Science, Engineering and Education Applications (ICCSEEA): <https://ieeexplore.ieee.org/document/9504117>
- [14] "A Hybrid Machine Learning Approach for Phishing Detection" by A. E. B. S. AlShrouf, M. Al-Nashash, and S. Aljawarneh, published in the Proceedings of the 2021 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT): <https://ieeexplore.ieee.org/document/9403408>
- [15] "Phishing Detection: A Machine Learning-Based Approach" by M. A. Aziz, R. H. Bijon, and M. N. Sultana, published in the Proceedings of the 2022 3rd International Conference on Computer, Communication and Computational Sciences (IC4S): <https://ieeexplore.ieee.org/document/9661601>





Impact Factor: 8.379



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  [ijircce@gmail.com](mailto:ijircce@gmail.com)



[www.ijircce.com](http://www.ijircce.com)

Scan to save the contact details