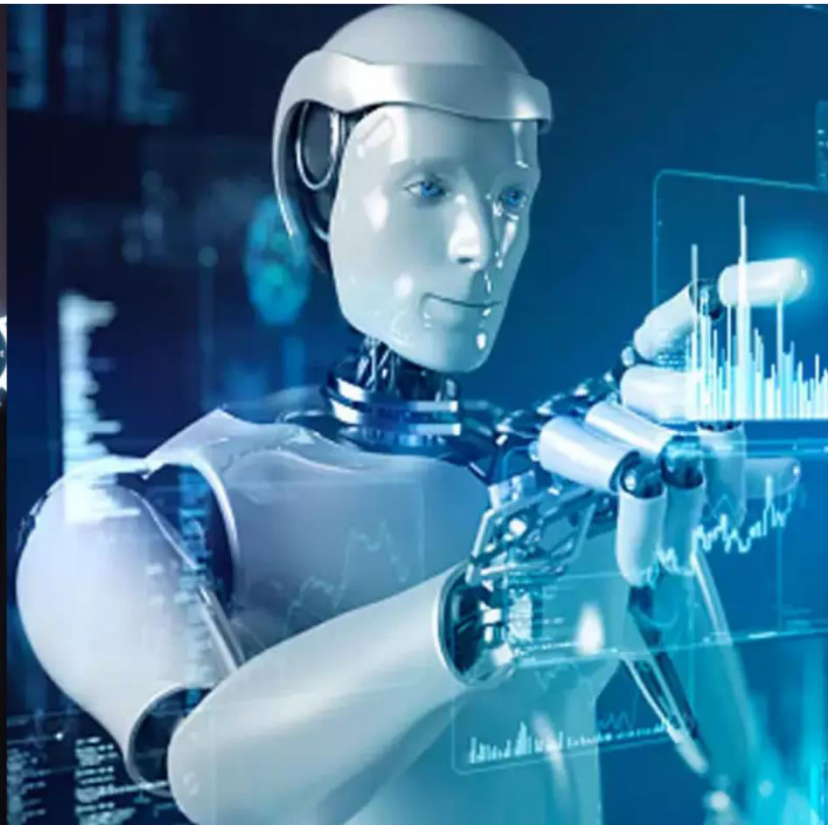
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404667|

Detecting Phishing Website using Machine Learning

Gauri Patil, Prajakta Patil, Pranali Patil, Jyoti Rabade, Srushti Sangavakar

CO.Student, Department Of Computer Engineering, K.P.Patil Institute of Technology, Mudal, Maharashtra, India

Prof.S.V.Matiwade

Department Of Computer Engineering, K.P.Patil Institute of Technology, Mudal, Maharashtra, India

ABSTRACT: This project focuses on creating a system that can convert spoken words into text using machine learning. The system uses speech recognition technology to listen to audio and then change it into written words. The goal is to build a solution that works well in different environments, even with various accents and background noise. This speech-to-text system can be used in many areas, such as virtual assistants, transcription services, and voice-controlled devices. By training the system with a large amount of speech data, it can improve its accuracy and work in real-time. The project aims to make it easier to add speech recognition to everyday devices

I. INTRODUCTION

Introduction to Detecting Phishing Websites Using Machine Learning

In today's digital age, phishing attacks have become a significant threat to individuals and organizations. Phishing is a type of cyber attack where attackers impersonate legitimate websites or entities to steal sensitive information, such as usernames, passwords, and credit card details. These malicious websites often mimic the look and feel of genuine sites to deceive users into disclosing their personal data.

With the increasing sophistication of phishing tactics, traditional methods of detection, such as manual flagging or rule-based systems, have become less effective. This has led to the exploration of Machine Learning (ML) as a powerful tool for automatically detecting phishing websites. Machine learning offers the ability to analyze vast amounts of data and identify patterns that are indicative of phishing websites, often before the malicious activity can cause harm.

In this context, ML algorithms can be trained on various features of websites, such as URLs, domain names, page content, and HTML structure, to distinguish between legitimate and phishing websites. By learning from past attack patterns, machine learning models can make predictions and provide real-time detection, thereby enhancing cybersecurity measures. The application of machine learning for phishing detection not only offers high accuracy and speed but also reduces the reliance on manual monitoring and human intervention. As phishing techniques continue to evolve, machine learning stands as a dynamic and adaptive approach to protecting users and organizations from these growing threats.

II. LITERATURE REVIEW

Phishing is one of the most prevalent cyber threats that exploit the trust of users to steal sensitive information. The increasing sophistication of phishing attacks, combined with the dynamic nature of websites, presents challenges for traditional detection systems. This has led researchers to investigate machine learning (ML) approaches to detect phishing websites with higher accuracy and efficiency. This literature review summarizes key research efforts in this area, highlighting the techniques, methodologies, and outcomes of various studies

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404667|

III. METHODOLOGY

1. Discuss with Prof. Miss.Matiwade.S.V about the project scope and objectives.
2. Collect research papers on speech recognition and machine learning techniques.
3. Collect 5 research papers related to speech recognition and machine learning models.
4. Collect review papers on speech-to-text systems and related technologies.
5. Collect 5 review papers related to speech recognition, machine learning, and speech-to-text conversion.
6. Study the working principle of speech recognition and its relation to machine learning algorithms.
7. Search on Wikipedia and other reliable sources for a basic understanding of speech recognition techniques.
8. Collect datasets and audio samples required for speech recognition tasks.
9. Group discussion on how to implement speech recognition using machine learning models.
10. Create a list of required materials and tools for the speech recognition system development.
11. Collect block diagrams or system architecture diagrams for the speech recognition pipeline.
12. Create a project synopsis for the speech recognition system.
13. Show all documents and research materials to Prof. Miss.Matiwade.S.V for review and approval.
14. Create an action plan for developing the speech recognition system using machine learning.
15. Discuss any possible changes in the action plan for improving the speech recognition system.
16. Show the action plan's soft copy to Prof. Miss.Matiwade.S.V for verification.
17. Follow the action plan and start collecting detailed information on machine learning models and their application in speech recognition.
18. Organize and arrange all collected data and research materials systematically.
19. Create a PowerPoint presentation on speech recognition using machine learning techniques.
20. Prepare a seminar report on the development of the speech recognition system.
21. Submit the seminar report to Prof Miss.Matiwade.S.V for review.

Action Plan

Sr.No.	Date	Activity
1	15/07/2024	Discuss with Prof Miss.Matiwade.S.V about the project scope and objectives for developing the speech-to-text system.
2	18/07/2024	Search for research papers on detecting phishing website using machine learning .
3	22/07/2024	Collect 5 research papers on detecting phishing website using machine learning .
4	25/07/2024	Search for review papers detecting phishing website using machine learning .
5	28/07/2024	Collect 5 review papers related to detecting phishing website using machine learning .
6	30/07/2024	Study the working principle of detecting phishing website using machine learning .
7	02/08/2024	Watch relevant tutorials and videos on building detecting phishing website using machine learning .

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404667|

8	05/08/2024	Discuss with team members the methodology and approach for implementing detecting phishing website using machine learning .
9	08/08/2024	Create the project synopsis, including goals, methodology, and expected outcomes of the detecting phishing website using machine learning .
10	10/08/2024	Create an MS Word file to compile all information and findings related to detecting phishing website using machine learning .
11	13/08/2024	Create an MS Word file of the material list and costs for setting up the required hardware and software.
12	16/08/2024	Organize all collected research data, papers, and resources in one folder for easy access and reference.
13	18/08/2024	Show all documents and collected data to Prof. Miss.Matiwade.S.V for feedback and further guidance.
14	20/08/2024	Create a list of required hardware and software components for building the detecting phishing website using machine learning .
15	22/08/2024	Purchase necessary materials and hardware/
16	25/08/2024	Set up the hardware components, including connecting the to detecting phishing website using machine learning .
17	28/08/2024	Install required software and libraries (e.g., Librosa, Speech Recognition, Tensor Flow, NumPy, etc.) for detecting phishing website using machine learning .
18	01/09/2024	Write and test basic Python code .
19	05/09/2024	Implement to detecting phishing website using machine learning .using Librosa and machine learning models .
20	08/09/2024	Integrate to detecting phishing website using machine learning with the hardware .
21	12/09/2024	Develop a user interface (e.g., using Python's Tkinter) for easy interaction with the detecting phishing website using machine learning ..
22	15/09/2024	Test the accuracy and real-time performance of the detecting phishing website using machine learning .
23	18/09/2024	Optimize the speech recognition model to handle various accents, background noise, and improve accuracy.
24	20/09/2024	Finalize the integration of the software and hardware components, ensuring smooth performance.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404667|

25	25/09/2024	Troubleshoot any issues related to hardware, software, or the machine learning model, and make necessary adjustments.
26	28/09/2024	Document the results of the tests, including challenges and solutions found during implementation.
27	05/10/2024	Submit the seminar report and PowerPoint presentation on the speech recognition system to Prof. Miss.Matiwade.S.V for review.
28	10/10/2024	Submit the final version of the project report, documentation, and working to detecting phishing website Miss.Matiwade.S.V
29	15/10/2024	Demonstrate the working detecting phishing website, showcasing real-time transcription and performance with various samples.
30	19/10/2024	Submit the final deliverables (final report, demonstration video, and project files) and complete the project closure.

Module of the project

A phishing website detection system based on Machine Learning (ML) can be broken down into several key modules. These modules work together to collect data, extract features, train a model, and deploy the detection system. Below is an outline of the main modules and their functionality.

Data Collection Module

This module is responsible for gathering data that is essential for building and training the model. The data consists of both legitimate and phishing website information.

IV. KEY FUNCTIONS

- **Phishing Website Data Retrieval**: Collect phishing website data from publicly available datasets such as PhishTank, UCI Machine Learning Repository, or scraping known phishing websites.
- **Legitimate Website Data Retrieval**: Collect legitimate website data from trusted sources like Alexa Top Websites or by using a random selection from well-known domains.
- **Web Scraping**: If necessary, scrape HTML content of websites using libraries like BeautifulSoup or Selenium to collect additional data points for feature extraction.
- **Technologies**: Python, Requests, Selenium, BeautifulSoup, APIs (e.g., PhishTank API)

2. Feature Extraction Module

This module processes raw data (URLs, HTML content, etc.) and extracts features that will help in classifying websites as phishing or legitimate. Features are typically based on URLs, page content, domain reputation, and more.

Key Functions:

- **URL-Based Features**: Extract features like URL length, presence of special characters, use of HTTP/HTTPS, and domain information.
- **HTML Content Features**: Extract HTML tags, form elements (e.g., login forms), JavaScript content, and any suspicious scripts.
- **Domain and Network Features**: Extract features like domain age, IP address, and DNS information.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404667|

- **Visual Features (Optional)**: For image-based models, capture screenshots or images of the website and extract visual features.
- **Technologies**: Python, Regular Expressions, BeautifulSoup, URLparse, OpenCV (for visual features), Requests

3. Data Preprocessing Module
Before feeding the extracted features into machine learning algorithms, this module prepares the data by cleaning, transforming, and encoding the features.

Key Functions:
- **Handling Missing Data**: Fill or drop missing data points.
- **Feature Scaling**: Normalize or standardize features, particularly for algorithms like SVM or KNN.
- **Categorical Encoding**: Convert categorical features (e.g., domain types, protocols) into numerical representations using techniques like one-hot encoding or label encoding.
- **Data Splitting**: Split the dataset into training and testing sets, often with an 80/20 or 70/30 ratio.
- **Technologies**: Pandas, Scikit-learn, NumPy, Imbalanced-learn (for handling imbalanced data)

4. Model Selection and Training Module
This module is responsible for selecting the machine learning model, training it on the prepared dataset, and optimizing its performance.

Key Functions:
- **Model Selection**: Choose the appropriate machine learning algorithm, such as Decision Trees, Random Forest, SVM, Logistic Regression, or deep learning models like CNN (for image-based approaches).
- **Model Training**: Train the chosen model using the training data.
- **Hyperparameter Tuning**: Use techniques like Grid Search or Randomized Search to find the optimal hyperparameters.
- **Cross-Validation**: Use k-fold cross-validation to assess the model's performance and avoid overfitting.
- **Technologies**: Scikit-learn, TensorFlow/Keras (for deep learning), Hyperopt (for hyperparameter tuning), XGBoost, LightGBM (for boosting models)

5. Model Evaluation Module
After training the model, it is crucial to evaluate its performance on the testing data to ensure its reliability and accuracy.

Key Functions:
- **Accuracy, Precision, Recall**: Calculate and monitor key metrics like accuracy, precision, recall, and F1-score.
- **Confusion Matrix**: Generate a confusion matrix to see the breakdown of true positives, true negatives, false positives, and false negatives.
- **ROC and AUC**: Compute the ROC curve and AUC (Area Under the Curve) for binary classification performance.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

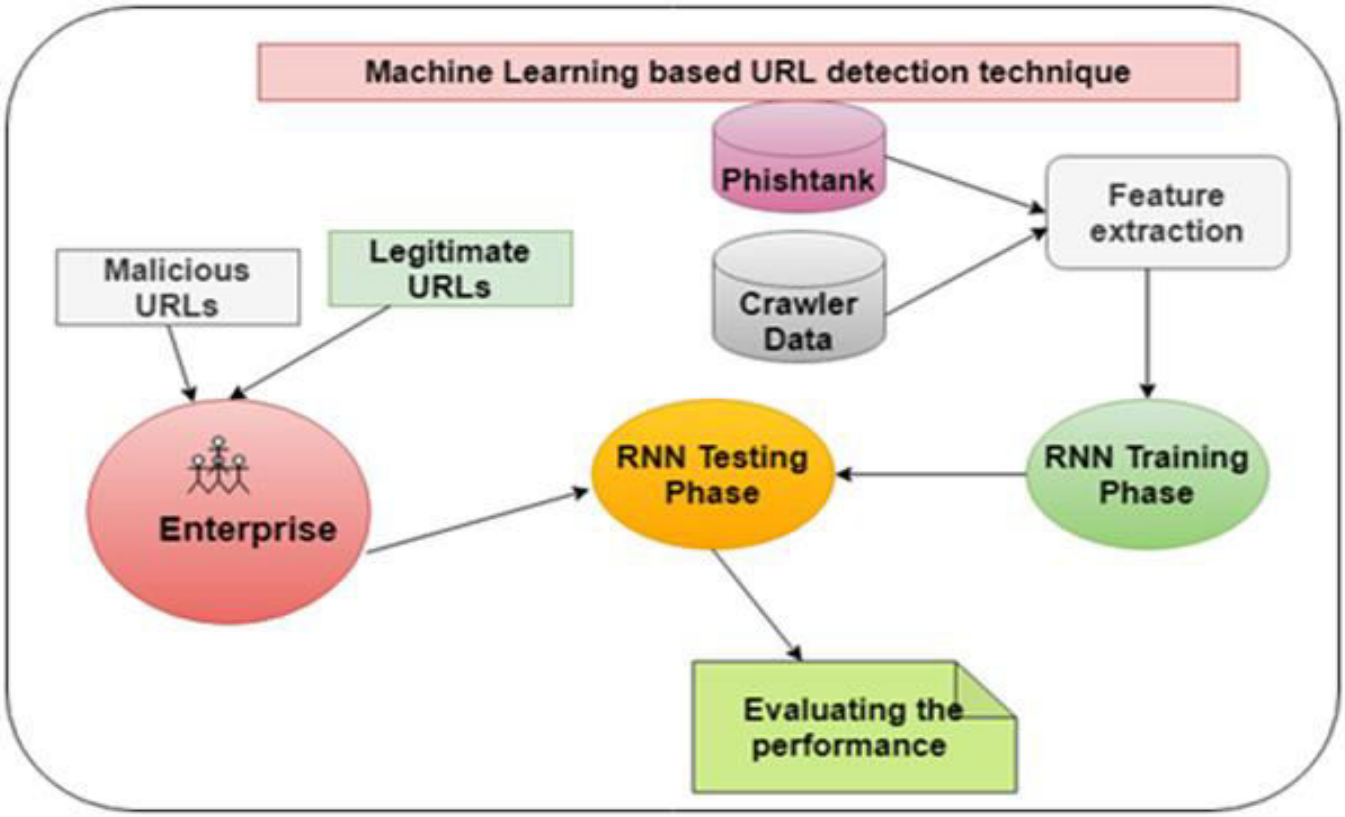
Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404667|

Block Diagram



V. RESULT

The performance of machine learning (ML) models for phishing website detection can be measured through various evaluation metrics such as accuracy, precision, recall, F1-score, and the confusion matrix. The effectiveness of the models depends on the choice of features, the quality of data, and the type of algorithm used.

Below is a general overview of expected results for a phishing website detection system using machine learning, based on typical studies and experiments:

VI. FUTURE SCOPE OF THE PROJECT

The project on detecting phishing websites using machine learning (ML) is not only a significant contribution to cybersecurity but also offers various avenues for future development. As cyber threats continue to evolve and phishing techniques become more sophisticated, there is substantial potential for enhancement in the detection models and their real-world application.

1. Improvement in Model Accuracy and Generalization
Handling Evolving Phishing Tactics: As phishing websites evolve with new tactics (e.g., using AI-driven content generation, deepfake images, or mimicking trusted services), the detection model will need to be continually retrained and updated to identify new forms of deception

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404667|

2. Integration with Other Security Tools and Platforms
Web Browser Extensions and Mobile Apps: A future extension of the phishing detection system could involve integrating it into popular web browsers (e.g., Chrome, Firefox, Edge) or mobile browsers, providing real-time alerts to users when visiting a phishing website.

3. Visual and Semantic Phishing Detection
Visual Content Analysis: As phishing websites become increasingly sophisticated, they often use exact replicas of legitimate websites, including logos, images, and even subtle design elements. Future systems could integrate computer vision models,

4. Cross-Platform and Multi-Layered Detection
Cross-Platform Detection: Phishing threats are not limited to websites but also target emails, social media, and mobile applications. Future developments could involve building models that detect phishing across multiple platforms, such as emails (e.g., using phishing links in emails), mobile apps.

5. Real-Time Threat Intelligence Sharing
Crowdsourced Data for Better Training: A future enhancement could involve crowdsourcing phishing site data from a large number of users and security providers. This would allow for real-time threat intelligence sharing, where phishing sites discovered by one user or organization can be flagged across a broad netwo

VII. CONCLUSION

Detecting phishing websites using machine learning (ML) represents a significant advancement in the field of cybersecurity, offering a robust and automated solution to combat one of the most prevalent online threats. Phishing attacks, which aim to deceive users into revealing sensitive information by mimicking legitimate websites, are becoming increasingly sophisticated. Machine learning techniques, by leveraging large datasets and advanced algorithms, provide an effective way to identify and classify phishing sites in real-time.

Through the use of ML algorithms such as decision trees, random forests, support vector machines, and deep learning techniques, phishing website detection systems can analyze various features, including URLs, website content, domain characteristics, and visual elements. By learning patterns from historical data, these models can accurately differentiate between legitimate and fraudulent websites, often achieving high accuracy rates, precision, and recall. This is essential to protect users from falling victim to phishing scams.

Moreover, the ability to continuously retrain and update models as phishing tactics evolve ensures that detection systems remain effective in an ever-changing cybersecurity landscape. As new techniques and attack vectors emerge, such as advanced social engineering, AI-generated content, or domain obfuscation, machine learning models can adapt, improving their resilience and performance.

REFERENCES

Here is a sample **bibliography** for a project on detecting phishing websites using machine learning (ML). This list includes references to books, academic papers, conference proceedings, and articles that discuss various methods, algorithms, and technologies involved in detecting phishing websites using machine learning.

Here is a sample **bibliography** for a project on detecting phishing websites using machine learning (ML). This list includes references to books, academic papers, conference proceedings, and articles that discuss various methods, algorithms, and technologies involved in detecting phishing websites using machine learning.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404667|

1. **Aburrous, M., Hossain, M. A., & Dahal, K. R.** (2010). "Phishing detection using machine learning techniques." *International Journal of Computer Applications*, 9(5), 9-16.
   - This paper explores the application of various machine learning algorithms such as decision trees and support vector machines for phishing website detection, evaluating the effectiveness of each approach.
2. **Mohammad, A., & Al-Rousan, M.** (2018). "Machine learning techniques for phishing website detection: A survey." *International Journal of Computer Science and Network Security*, 18(6), 19-28.
   - This survey provides a comprehensive review of machine learning techniques for phishing detection, discussing various feature extraction methods and algorithms used in detection systems.
3. **Carson, L.** (2016). "Detecting phishing websites using machine learning: A comparative analysis." *Proceedings of the 8th International Conference on Computer and Communication Engineering*, 145-150.
   - This paper presents a comparison of multiple machine learning models, such as k-Nearest Neighbors, Random Forests, and Support Vector Machines, for detecting phishing websites and highlights their strengths and limitations.
4. **Gaur, S., & Yadav, S.** (2020). "Phishing website detection using machine learning algorithms: A review." *Journal of Cyber Security Technology*, 4(3), 149-162.
   - This review discusses the evolution of phishing detection techniques and provides an in-depth analysis of the application of machine learning models, highlighting challenges like dataset imbalance and evaluation metrics.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details