Slide 1:

Detecting Malicious URLS
Detecting Malicious URLs Using Machine Learning Techniques
RAJEEV GANDHI MEMORIAL COLLEGE OF ENGINEERING & TECHNOLOGY
Approved by AICTE - New Delhi Affiliated to JNTU Anantapuramu
nandyal - 518501, andhra pradesh
(Autonomous)

Observations by

S. Abdul Hameed – 22091A3201
S. Arshiya Parveen – 22091A3210
Under Mentorship of

Dr. B. Bhaskara Rao

Slide 2:

Overview

Slide 3:

Introduction
The digital world is advancing rapidly, with increased online activity.
Cyberattack risks are rising due to attackers' inventive techniques.
Malicious URLs are a critical attack vector, used to extract information and trick users.
This review examines machine learning (ML) techniques for detecting malicious URLs.
It addresses the lack of research on detecting malicious Arabic websites

3

Slide 4:

Why it is Important?
Malicious URLs can lead to system hacks and sensitive data breaches.
They cause billions of dollars in losses each year.
Traditional blacklisting methods have limitations and can be bypassed.
ML offers a promising solution by learning from experience and improving self-learning without human intervention.
Securing websites is crucial for online activities like e-commerce, business, social networking, and banking

4

Slide 5:

Datasets used
Arabic studies rely on custom-built datasets due to the lack of public Arabic datasets
5

Slide 6:

Feature Extraction
6

Slide 7:

Machine Learning Models Used
Algorithms achieving high accuracy (>=99%): CNN, XGBoost, LSTM, SVM, CW, Majority Voting Classifier, RF, K-means, Ara-means, DT, NB.
Frequently used algorithms with good performance: SVM, RF, DT, NB, and LR.
Ensemble techniques often provide high accuracy.
Algorithms with lower performance: BN, NN, DBN.

7

Slide 8:

Model Training and Evaluation
Studies reviewed used supervised, unsupervised, and semi-supervised learning.
Most studies used binary classification (malicious or benign).
Evaluation metrics varied across studies, but accuracy, precision, and recall were common.
K-fold cross-validation was used in some studies
Accuracy scores varied from model to model by achieving 90-99.7%

8

Slide 9:

Applications
Malicious URL detection is crucial for various applications:

Web security.
Email security.
Network security.
Protecting online transactions and user data.
Combating phishing, spam, and malware attacks.

9

Slide 10:

Findings and Results
Lexical features are frequently used in both Arabic and English content analysis.
Content-based features are more common in Arabic studies.
Network-based features are not used in Arabic content analysis.
The highest accuracy in English studies was 99.98% (CNN).
The highest accuracy in Arabic studies was 99.521% (DT).
Key network-based features include domain country code, domain update time, and contract expiration time.

10

Slide 11:

Challenges and Our Extensions

Slide 12:

Conclusion
Lexical features are frequently used, and SVM, RF, CNN, and XGBoost are effective algorithms.
Future research should focus on addressing challenges like dataset size and feature selection and detector sustainability.
Further research is needed to enhance the accuracy and robustness of malicious URL detection techniques.

Slide 13:

Thank you