DETECTION OF MALICIOUS WEBLINKS THROUGH DOMAIN FEATURE EXTRACTION

INTRODUCTION

- Phishing attacks commonly exploit user trust by mimicking legitimate websites using deceptive URLs or weblinks. These weblinks often contain suspicious keywords, recently registered domains, uncommon TLDs, or obfuscated structures designed to bypass traditional security filters.
- This project presents a Chrome browser extension that performs real-time analysis of weblinks to detect phishing attempts. The system extracts structural and contextual features from each URL, such as domain age, special characters, subdomain levels, and lookalike patterns. It classifies each URL as either phishing or legitimate, provides a confidence score, and categorizes it into domains such as financial, e-commerce, or social platforms.
- The tool offers visual warnings, explanations, and historical tracking of phishing detections, enhancing user awareness and enabling a safer browsing experience.

PROBLEM STATEMENT

Despite the prevalence of phishing attacks, current browser defenses are reactive, not proactive — often relying on static blocklists or delayed reputation updates.

Many phishing URLs bypass these defenses by:

- Using newly registered domains with no prior history.
- Obfuscating intent through special characters, excessive subdomains, or IP-based URLs.
- Mimicking trusted services with lookalike patterns (e.g., paypa1.com, micr0soft-login.xyz).

Furthermore:

- Most detection systems do not categorize weblinks by domain sector, missing insights into which industries are targeted.
- Real-time alerts, feature-based analysis, and user-side customization are typically absent in available solutions.
- Users lack transparent feedback on why a link is flagged and what makes it risky.

This project addresses these gaps by developing a client-side detection and classification tool that empowers users with actionable insights and timely protection.

LIMITATIONS

1.Blacklist Dependence:

Could only detect previously known phishing URLs; failed against new or evolving attacks.

2. Static Heuristic Rules:

Used fixed rules that lacked adaptability, often resulting in high false positives or negatives.

3. Lack of Real-Time Detection:

Most systems required manual URL input and did not offer real-time browser protection.

4. Limited Dataset Usage:

Many models were trained on small datasets, reducing their effectiveness in real-world scenarios.

SYSTEM ARCHITECTURE

1. User Input Layer (Frontend Interface)

- Component: popup.html, popup.js
- Purpose: Provides an interface for the user to:
  - Input a URL manually or analyze the current browser tab
  - View detection results, settings, and statistics
- UI Tabs:
  - Analyzer Tab: Manual/current URL input
  - Settings Tab: Options like domain age check, TLD alerts, sensitivity
  - Statistics Tab: Shows phishing trends by category/domain
  - History Tab: Lists past detections with time and score

2. Background Logic Layer (Main Engine)

- Component: background.js
- Purpose: Central logic layer that handles:
- Listening to messages from the popup (URL requests)
- Triggering analysis, scoring, classification, and response
- Interfacing with Chrome APIs (tabs, storage, runtime)

3. Feature Extraction Engine

- Component: background.js + utils.js
- Process:
  - Normalize the URL (ensure protocol, extract domain/path)
  - Extract features such as:
  - Domain length, path length,Number of dots, hyphens, digits, equal signs, subdomains , Domain registration age (mocked), Use of IP address or @

4. Scoring and Detection Engine

- Component: background.js
- Apply heuristic-based logic (rule-based model)
- Assign score to each feature (weighted logic)
- Adjust score based on sensitivity level (user-defined)
- Cap score at 0.95, then compare with threshold
- Logic:
- score = ∑(featureWeight × multiplier)
- threshold = 0.3 - ((sensitivityLevel - 3) × 0.05)
- If score > threshold → phishing, else legitimate

5. Result Handling and Classification

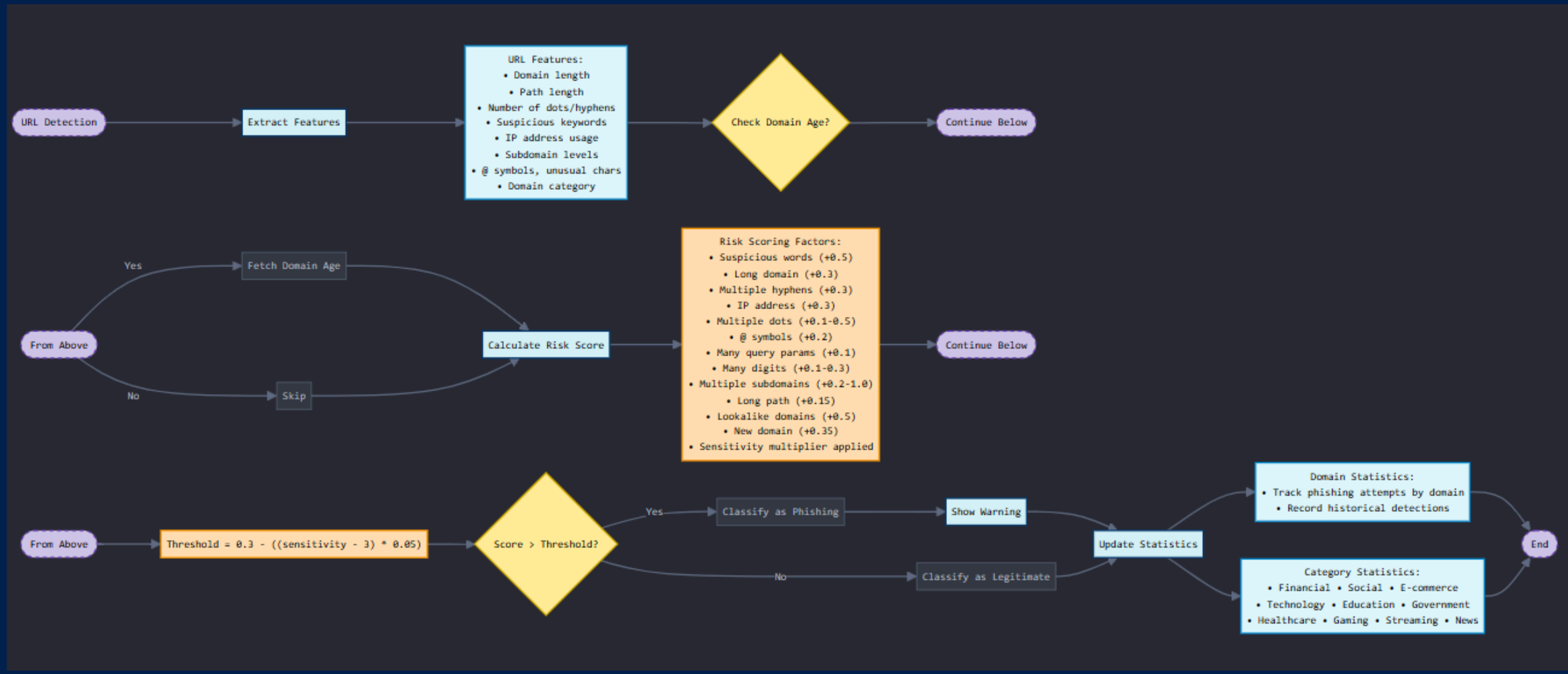
- Component: background.js → returns to popup.js
- Result Returned:
    - Final classification: Phishing or Legitimate

- Confidence score
- Risk explanation (why it was flagged)
- Domain category (e.g., finance, social)
- Domain age status
- TLD warning (if present)
- Phishing history for the domain

6. Content Script Action (Page Injection)

- If the analyzed URL is detected as phishing, inject a warning overlay directly into the webpage
- Prevents user interaction unless they choose to proceed

WORKFLOW


URL Detection
Extract Features
URL Features:
• Domain length
• Path length
• Number of dots/hyphens
• Suspicious keywords
• IP address usage
• Subdomain levels
• @ symbols, unusual chars
• Domain category
Check Domain Age?
Continue Below
Yes
Fetch Domain Age
From Above
No
Skip
Calculate Risk Score
Risk Scoring Factors:
• Suspicious words (+0.5)
• Long domain (+0.3)
• Multiple hyphens (+0.3)
• IP address (+0.3)
• Multiple dots (+0.1-0.5)
• @ symbols (+0.2)
• Many query params (+0.1)
• Many digits (+0.1-0.3)
• Multiple subdomains (+0.2-1.0)
• Long path (+0.15)
• Lookalike domains (+0.5)
• New domain (+0.35)
• Sensitivity multiplier applied
Continue Below
From Above
Threshold = 0.3 - ((sensitivity - 3) * 0.05)
Score > Threshold?
Yes
Classify as Phishing
Show Warning
No
Classify as Legitimate
Update Statistics
Domain Statistics:
• Track phishing attempts by domain
• Record historical detections
Category Statistics:
• Financial • Social • E-commerce
• Technology • Education • Government
• Healthcare • Gaming • Streaming • News
End

PROPOSED SYSTEM

MODULE 1 : PREPROCEESING WITH FEATURE EXTRACTION

Objective: Prepare the input data (URLs) and extract meaningful features for analysis.
Steps Involved:

- Input Handling:
- URLs are fetched and cleaned — removing irrelevant characters or whitespaces.
- Normalization:
- Standardize the format of URLs to avoid inconsistencies (e.g., lowercase conversion, removing http://, etc.).

- Feature Extraction (FE):
- This is the most crucial part. The system analyzes the structure of the URL and extracts multiple features such as:
  - Presence of @ symbol
  - Number of subdomains
  - Number of dots (.)
  - Use of IP address instead of a domain name
  - Presence of suspicious keywords (like login, secure, verify, etc.)
  - Length of the URL
  - Use of HTTPS or not
  - Presence of suspicious characters like -, =, etc.
- Confidence Scoring:
- Based on extracted features, each URL is given a base risk score (initial confidence level). This score becomes the input for the next module.

MODULE 2 : APPLICATION OF ML ALGORITMS

Objective: Analyze the extracted features using machine learning logic to refine the confidence score and detect phishing patterns.

Steps Involved:

- Rule-Based Logic (Mimics ML):
- Even if not a trained ML model, you apply a weighted logic similar to a decision-tree approach:
  - If certain conditions are met (e.g., many dots, presence of suspicious words), the score is incremented.
  - If multiple red flags are detected, the score accumulates accordingly.
  - Mathematical logic like score += 0.1 * min(dots, 5) is used.

- Multipliers:
- Some features (like presence of “@” or IP address) apply multipliers to increase the phishing likelihood score exponentially.
- Thresholding:
- A final threshold (e.g., 0.6 or 60%) is set to distinguish between legitimate and phishing URLs.

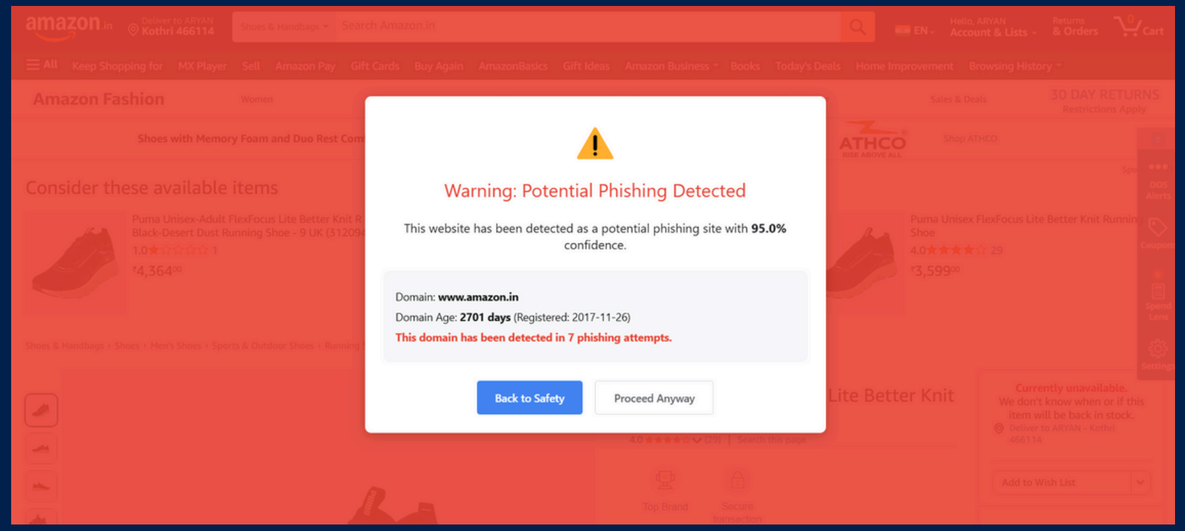
MODULE 3 : FINAL URL CLASSIFICATION AND RESULT DISPLAY

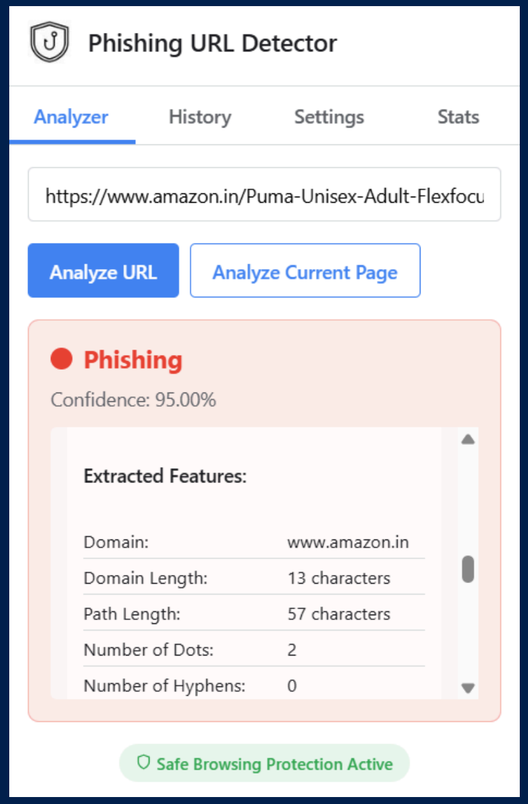
Objective: Use the calculated score to classify the URL and notify the user.
Steps Involved:

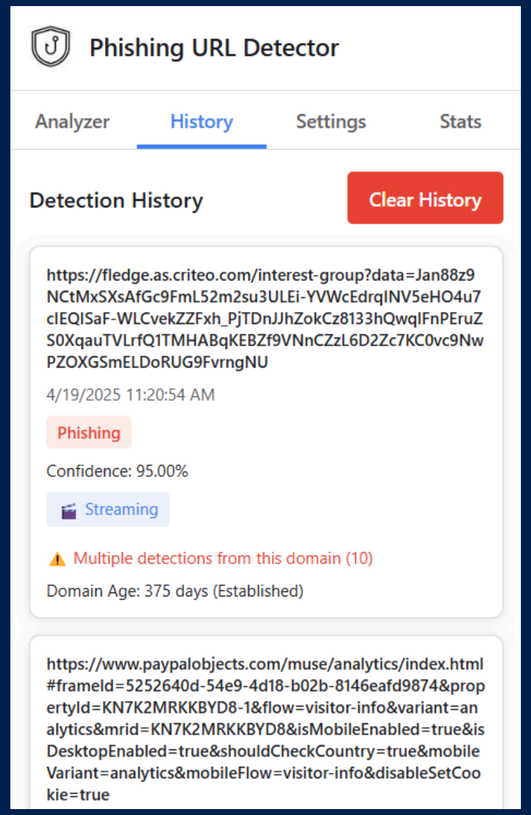
- Score Interpretation:
- Based on the total confidence score:
  - Score < 0.4 → Safe
  - Score between 0.4–0.6 → Suspicious
  - Score > 0.6 → Phishing
- Classification Output:

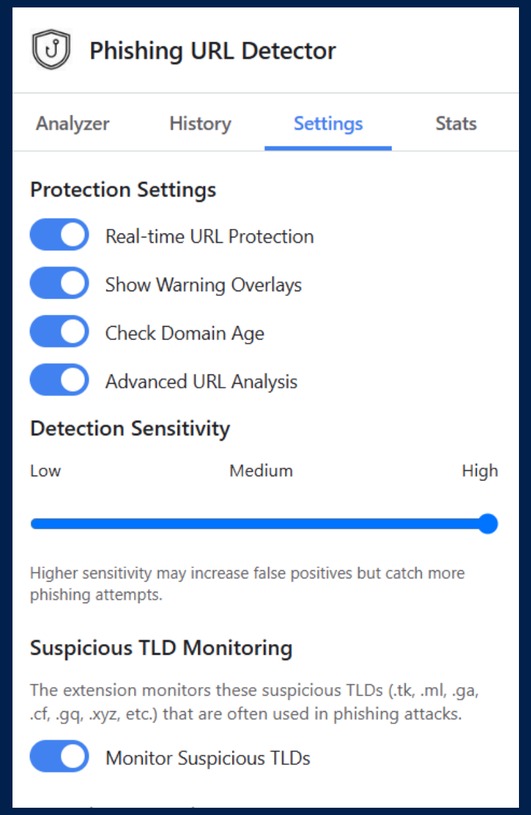
- The result is dynamically shown in the Chrome extension popup using clear labels and colors (e.g., green for safe, red for phishing).
- User Alert:
- Users are alerted about phishing risks and advised to avoid unsafe links.
- Continuous Use:
- Each time a URL is clicked or inspected, the same pipeline is applied.

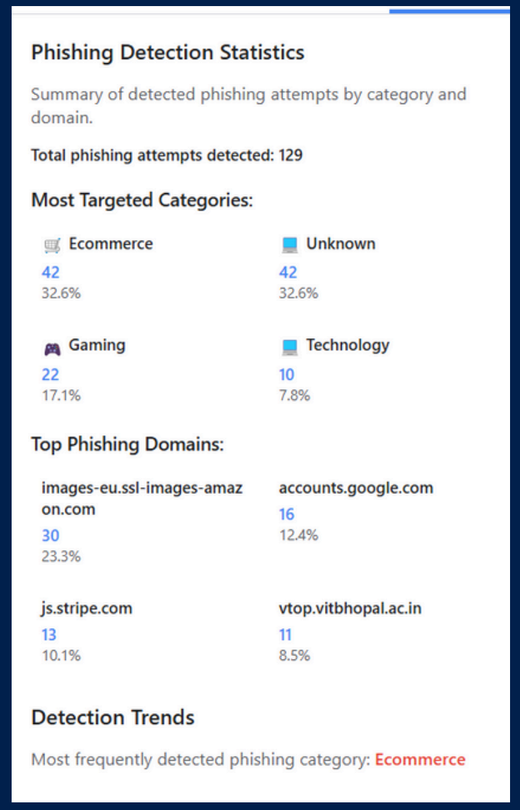
OUTPUT


Warning: Potential Phishing Detected
This website has been detected as a potential phishing site with 95.0% confidence.
Domain: www.amazon.in
Domain Age: 2701 days (Registered: 2017-11-26)
This domain has been detected in 7 phishing attempts.
Back to Safety
Proceed Anyway


Phishing URL Detector
Analyzer
History
Settings
Stats
https://www.amazon.in/Puma-Unisex-Adult-Flexfocu
Analyze URL
Analyze Current Page
Phishing
Confidence: 95.00%
Extracted Features:
Domain: www.amazon.in
Domain Length: 13 characters
Path Length: 57 characters
Number of Dots: 2
Number of Hyphens: 0
Safe Browsing Protection Active


Phishing URL Detector
Analyzer
History
Settings
Stats
Detection History
Clear History
https://fledge.as.criteo.com/interest-group?data=Jan88z9NCtMxSXsAfGc9FmL52m2su3ULEi-YVWcEdrqINV5eHO4u7cIEQISaF-WLCvekZZFxh_PjTDnJJhZokCz8133hQwqIFnPEruZS0XqauTVLrfQ1TMHABqKEBZf9VNnCZzL6D2Zc7KC0vc9NwPZOXGSmELDoRUG9FvrngNU
4/19/2025 11:20:54 AM
Phishing
Confidence: 95.00%
Streaming
Multiple detections from this domain (10)
Domain Age: 375 days (Established)
https://www.paypalobjects.com/muse/analytics/index.html#frameId=5252640d-54e9-4d18-b02b-8146eafd9874&propertyId=KN7K2MRKKBYD8-1&flow=visitor-info&variant=analytics&mrid=KN7K2MRKKBYD8&isMobileEnabled=true&isDesktopEnabled=true&shouldCheckCountry=true&mobileVariant=analytics&mobileFlow=visitor-info&disableSetCookie=true


Phishing URL Detector
Analyzer
History
Settings
Stats
Protection Settings
Real-time URL Protection
Show Warning Overlays
Check Domain Age
Advanced URL Analysis
Detection Sensitivity
Low
Medium
High
Higher sensitivity may increase false positives but catch more phishing attempts.
Suspicious TLD Monitoring
The extension monitors these suspicious TLDs (.tk, .ml, .ga, .cf, .gq, .xyz, etc.) that are often used in phishing attacks.
Monitor Suspicious TLDs


Phishing Detection Statistics
Summary of detected phishing attempts by category and domain.
Total phishing attempts detected: 129
Most Targeted Categories:
Ecommerce
42
32.6%
Unknown
42
32.6%
Gaming
22
17.1%
Technology
10
7.8%
Top Phishing Domains:
images-eu.ssl-images-amazon.com
30
23.3%
accounts.google.com
16
12.4%
js.stripe.com
13
10.1%
vtop.vitbhopal.ac.in
11
8.5%
Detection Trends
Most frequently detected phishing category: Ecommerce

REFERENCE

MOHAMMAD, R. M., THABTAH, F., & MCCLUSKEY, L. (2014).
INTELLIGENT RULE-BASED PHISHING WEBSITES CLASSIFICATION.
IET INFORMATION SECURITY, 8(3), 153–160.
DOI: 10.1049/IET-IFS.2013.0202
✅THIS PAPER SUPPORTS THE IDEA OF USING ENGINEERED FEATURES + RULE SCORING FOR PHISHING CLASSIFICATION WITHOUT COMPLEX ML MODELS.

SALVI SIDDHI RAVINDRA, SHAH JUHI SANJAY, SHAIKH NAUSHEENBANU AHMED GULZAR, KHODKE PALLAVI (2021).
PHISHING WEBSITE DETECTION BASED ON URL.
INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN COMPUTER SCIENCE, ENGINEERING AND INFORMATION TECHNOLOGY (IJSRCSEIT)
DOI: 10.32628/CSEIT2173124
✅ THIS WORK PRESENTS A SUPERVISED LEARNING MODEL USING URL FEATURES, ALIGNING CLOSELY WITH THE SAME FEATURE LOGIC USED IN OUR BACKEND.

Thank You