PHÁT HIỆN URL LỪA ĐẢO SỬ DỤNG KỸ THUẬT HỌC CỘNG TÁC
DETECTING FRAUDULENT URLS USING COLLABORATIVE LEARNING

Nguyễn Hoàng Khánh Tuấn[1], Nguyễn Hoàng Long[2], Huỳnh Chí Văn[3]

[1]Khoa Công nghệ thông tin, thành phố Hồ Chí Minh, Việt Nam, tuannhk20@uef.edu.vn
[2]Khoa Công nghệ thông tin, thành phố Hồ Chí Minh, Việt Nam, longnh220@uef.edu.vn
[3]Khoa Công nghệ thông tin, thành phố Hồ Chí Minh, Việt Nam, vanhc20@uef.edu.vn

***Tóm tắt**: Tấn công lừa đảo đang là mối đe dọa trong sự phát triển của thời đại công nghệ 4.0. Các cuộc tấn công mạng ngày càng trở nên tinh vi, đặt ra rất nhiều thách thức cho các nhà an ninh mạng. Trong bài báo này, chúng tôi đề xuất phương phát hiện URL giả nạo bằng cách sử dụng kỹ thuật học hợp tác, sau đó chúng tôi xây dựng công cụ kiểm tra độ an toàn của một URL. Chúng tôi thực nghiệm trên bộ dữ liệu Kaggle, kết quả cho thấy thuật toán Gradient Boosting Classifier (GBC) có kết quả cao nhất với độ chính xác lên đến 97.4%.*

***Từ khóa**: Học cộng tác, tấn công mạng, URL lừa đảo.*

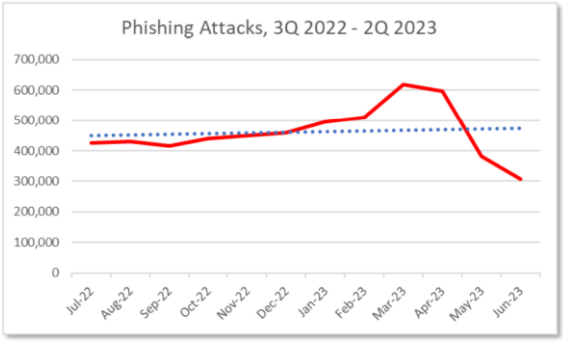
***Abstract**: Phishing attacks are a threat in the development of the 4.0 technology era. Cyber attacks are becoming increasingly sophisticated, posing many challenges for cyber security researchers. In this paper, we propose a method to detect fake URLs using collaborative learning techniques, then, we built a tool to check the safety of a URL. We experimented on the Kaggle data set, and the results showed that the Gradient Boosting Classifier (GBC) algorithm had the highest results with an accuracy of up to 97.4%.*

***Keywords**: Collaborative learning, Cybersecurity, machine learning, URL malicious*

1. Giới thiệu

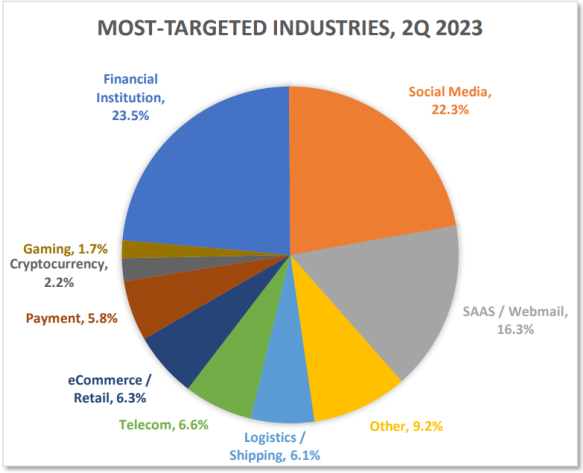
Ngày nay, nhu cầu sử dụng Internet đang ngày càng tăng cao, khi đó, nguy cơ dẫn đến các đợt tấn công trên không giang mạng là rất lớn. Hầu hết các kiểu tấn công lừa đảo trên không gian đều sử dụng URL lừa đảo. URL lừa đảo là các liên kết đến các trang web hoặc trang web được thiết kế trông giống như các trang web hợp pháp nhưng trên thực tế, chúng là các trang web độc hại do tội phạm mạng tạo.

Theo thống kê của Anti-Phishing Working Group (APWG) cho thấy trong 30,047 địa chỉ thì đã có 15,022 địa chỉ độc hại và 15,025 địa chỉ hợp lệ, dữ liệu đã được phân phối nhằm huấn luyện và kiểm tra dữ liệu. Dựa trên thống kê trong quý 2 năm 2023 của AWPG đã tổng hợp 1.286.208 cuộc tấn công lừa đảo. Đây là tổng số hàng quý cao thứ ba mà APWG từng ghi nhận. Tuy nhiên, đến đầu tháng 3 năm 2023 lừa đảo có xu hướng giảm *[1]* .



***Hình 1.** Biểu đồ thống kê tại APWG*

Trong đó, lĩnh vực tài chính luôn là lĩnh vực bị khai thác tấn công nhiều nhất với 23,5% tổng số cuộc tấn công lừa đảo *[1]* .



Hình 2. *Biểu đồ thống kê các lĩnh vực bị tấn công*

Mô hình học cộng tác được áp dụng để tận dụng sự thông tin từ nhiều nguồn khác nhau. Thay vì dựa chỉ vào dữ liệu cá nhân của từng người sử dụng, nghiên cứu này kết hợp thông tin từ cộng đồng người dùng, các dịch vụ bảo mật mạng và các nguồn thông tin khác nhau để xây dựng một hệ thống phòng ngừa mạnh mẽ. Phần còn lại chúng tôi trình bày như sau: Phần 2 chúng tôi trình bày kết quả khảo sát của các công trình nghiên cứu sử dụng máy học để phát hiện URL độc hại. Mô hình đề xuất chúng tôi trình bày và phân tích trong phần 3. Cuối cùng chúng tôi trình bày các kết quả đạt được của các thuật toán và hướng nghiên cứu tiếp theo trong tương lai.

2. Các công trình liên quan

Trong nhiều năm qua, đã có rất nhiều phương pháp được các nhà khoa học, nghiên cứu đề xuất nhằm mục đích phát hiện URL giả mạo có độ chính xác cũng khá khả quan.

Taeri Kim và các cộng sự (2022) *[2]* đã đề xuất một phương pháp suy luận dựa trên mạng để phát hiện chính xác các URL lừa đảo được ngụy trang bằng các mẫu hợp pháp. Phương pháp này được thiết kế để chống lại việc né tránh, có nghĩa là một URL lừa đảo vẫn sẽ được xác định là lừa đảo ngay cả khi nó đã được sửa đổi để trông hợp pháp. Kết quả thực nghiệm với khoảng 500000 URL phương pháp được đề xuất vượt trội hơn các phương pháp hiện đại nhất trong nhiều thí nghiệm khác nhau, đạt điểm F-1 là 89,1% so với 84% cho phương pháp dựa trên tính năng tốt nhất.

Shouq Alnemari và cộng sự (2023) *[3]* đã so sánh 4 mô hình nghiên cứu hiệu quả của công việc sử dụng máy học để phát hiện các tên miền lừa đảo, nó cũng so sánh các mô hình này được phát triển bằng các sử dụng mạng thần kinh nhân tạo (ANN), máy vectơ hỗ trợ (SVM), cây quyết định (DT) và kỹ thuật rừng ngẫu nhiên (RF). Hơn nữa, tập dữ liệu miền lừa đảo UCI của bộ định vị tài nguyên thống nhất (URL) được sử dụng làm điểm chuẩn để đánh giá các mô hình, bộ dữ liệu được lấy từ UCI phishing Websites. Kết quả nghiên cứu cho thấy đối với mỗi loại thuật toán có đo chính xác khác nhau: SVM: 94.66%, ANN: 95.5%, RF: 97.3%, DT: 96.3%.

Tao Feng và các cộng sự (2020) *[4]* khám phá phương pháp học sâu và phát triển bốn mô hình RNN (Mạng lưới thần kinh tuần hoàn) chỉ dựa trên các tính năng từ vựng của URL để phát hiện các cuộc tấn công lừa đảo. Sử dụng một tập dữ liệu gồm 1,5 triệu URL, nghiên cứu cho thấy các mô hình RNN được đề xuất đạt độ chính xác phát hiện vượt quá 96% mà không cần trích xuất tính năng thủ công. Nghiên cứu của tác giả sử dụng mô hình chuỗi URL và mã hóa 16 chiều dựa trên word2vec, mỗi mô hình sẽ dịch 1 ký tự của URL duy nhất từ vi sang vector 16 chiều ei. Với mỗi URL được dịch hoặc mã hóa sẽ được đưa vào lớp dầu của mạng

LSTM hoặc GRU dưới dạng chuỗi vector. Bộ dữ liệu tác giả đã xây dựng tập dữ từ kho lưu trữ phishtank. Tác giả đã sử dụng phương pháp học sâu và xây dựng 4 mô hình RNN chỉ sử dụng các đặc điểm từ vựng của URL để phát hiện các cuộc tấn công lừa đảo với độ chính xác hơn 96% và tác giả đã phát tiển kỹ thuật trực quan hóa độc đáo RTOD và thuật toán phát hiện trang phục quảng cáo IP dựa trên trạng thái RNN.

Nhìn chung, các phương pháp sử dụng ,áy học để phát hiện URL giả mạo đã đạt được những kết quả đáng kể.

3. Phương pháp đề xuất

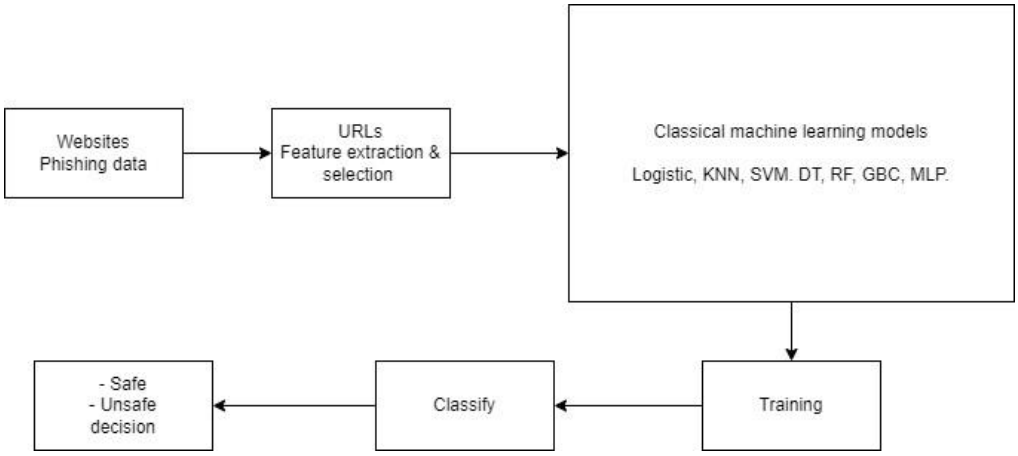
3.1. Mô hình đề xuất

Để cải thiện độ chính xác của việc phát hiện URL lừa đảo, chúng tôi đề xuất phương pháp sử dụng học hợp tác để phát hiện URL giả mạo:

(1) xử lý dữ liệu: chúng tôi tiến hành loại bỏ các cột đặc trưng không cần thiết và nhán dãn giá trị 1 (lành tính), -1 (lừa đảo).

(2) lựa chọn mô hình: chúng tôi sử dụng các mô hình máy học như RF, DT, GBC để thực nghiệm độ chính xác, sau đó chọn ra mô hình tối ưu nhất với độ chính xác cao nhất.

(3): kiểm tra kết quả: chúng tôi xây dựng công cụ để đánh giá một URL là lành tính hay lừa đảo.



Hình 3. *Mô hình thực nghiệm*

Trong nghiên cứu này, chúng tôi sử dụng 7 thuật toán bao gồm: Logistic Regression (LR) *[5]* , K-Nearest Neighbors (KNN) *[6]* , Support Vector Machine (SVM) *[7]* , Decision Tree (DT) *[8]* , Random Forest (RD) *[9]*, Gradient Boosting Classifier (GBC) *[10]* , Multi-layer Perceptron (MLP) *[11]* . Việc đánh giá dựa vào các chỉ số như:

Accuracy (ACC) tính toán trực tiếp bằng cách chia số lượng dự đoán đúng cho số lượng tất cả các dựu đoán.

$$ACC = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Precision là tỷ lệ các mẫu có liên quan trong số tất cả các mẫu được dự đoán thuộc về một lớp nhất định.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Giá trị Recall được định nghĩa là tỷ lệ các mẫu được dự đoán thuộc về một lớp so với tất cả cá mẫu thực sự thuộc về lớp đó.

$$Recall = \frac{TP}{TP + FN}$$

Chỉ số phân loại:

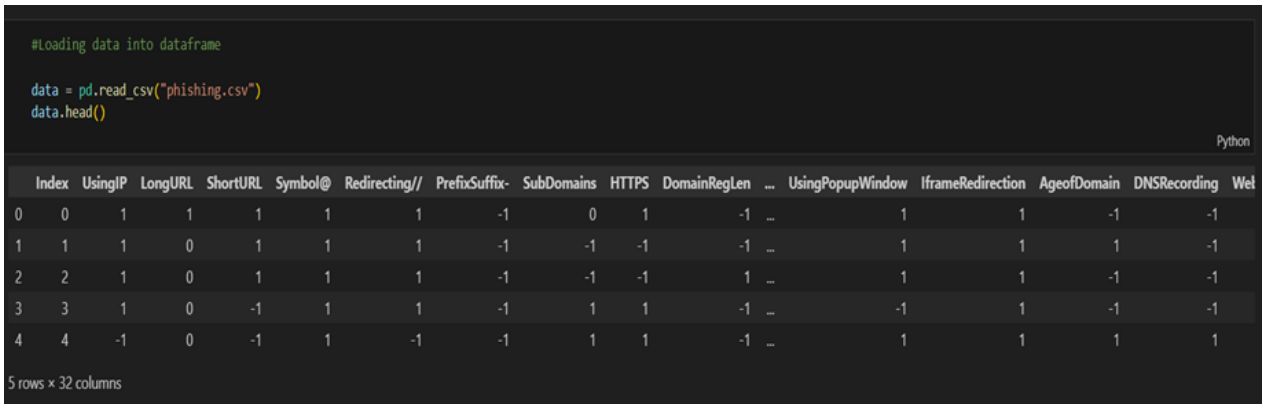
- True Positive (TP): Tổng số trường hợp dự báo khớp Positive.
- True Negative (TN): Tổng số trường hợp dự báo khớp Negative.
- False Positive (FP): Tổng số trường hợp dự báo các quan sát thuộc nhãn Negative thành trường hợp Positive.
- False Negative (FN): Tổng số trường hợp dự báo các quan sát thuộc nhãn Positive thành trường hợp Negative.

3.2. Bộ dữ liệu

Chúng tôi sử dụng bộ dữ liệu URL được cung cấp bởi Kaggle *[12]* . Bao gồm 11000 URL. Mỗi mẫu dữ liệu đều có 30 tham số và một lớp đối tượng được gán nhãn là 1 (lành tính) hoặc -1 (Độc hại). Bộ dữ liệu bao gồm 30 đặc trưng và được chia thành 4 nhóm:

Bảng 1: *Các nhóm và thành phần của bọ dữ liệu URL được cung cấp bởi Kaggle*

Nhóm	Thành phần
Liên quan đến URL	UsingIP, LongURL, ShortURL, Symbol@, Redirecting//, PrefixSuffix-, SubDomains, HTTPS, DomainRegLen, Favicon, NonStdPort, HTTPSDomainURL, RequestURL, AnchorURL, LinksInScriptTags, ServerFormHandler, AbnormalURL, WebsiteForwarding
Liên quan đến trình duyệt và giao diện người dùng	StatusBarCust, DisableRightClick, UsingPopupWindow, IframeRedirection
Liên quan đến thông tin về domain và website	AgeofDomain, DNSRecording, WebsiteTraffic, PageRank, GoogleIndex, LinksPointingToPage, StatsReport
Liên quan đến Email và giao tiếp	InfoEmail



```
#Loading data into dataframe

data = pd.read_csv("phishing.csv")
data.head()
```

Python

	Index	UsingIP	LongURL	ShortURL	Symbol@	Redirecting//	PrefixSuffix-	SubDomains	HTTPS	DomainRegLen	...	UsingPopupWindow	IframeRedirection	AgeofDomain	DNSRecording	Wel
0	0	1	1	1	1	1	-1	0	1	-1	...	1	1	-1	-1	
1	1	1	0	1	1	1	-1	-1	-1	-1	...	1	1	1	-1	
2	2	1	0	1	1	1	-1	-1	-1	1	...	1	1	-1	-1	
3	3	1	0	-1	1	1	-1	1	1	-1	...	-1	1	-1	-1	
4	4	-1	0	-1	1	-1	-1	1	1	-1	...	1	1	1	1	

5 rows × 32 columns

Hình 3.2: *Thông tin của bộ dữ liệu được lấy ở Kaggle*

3.3. Kết quả triển khai

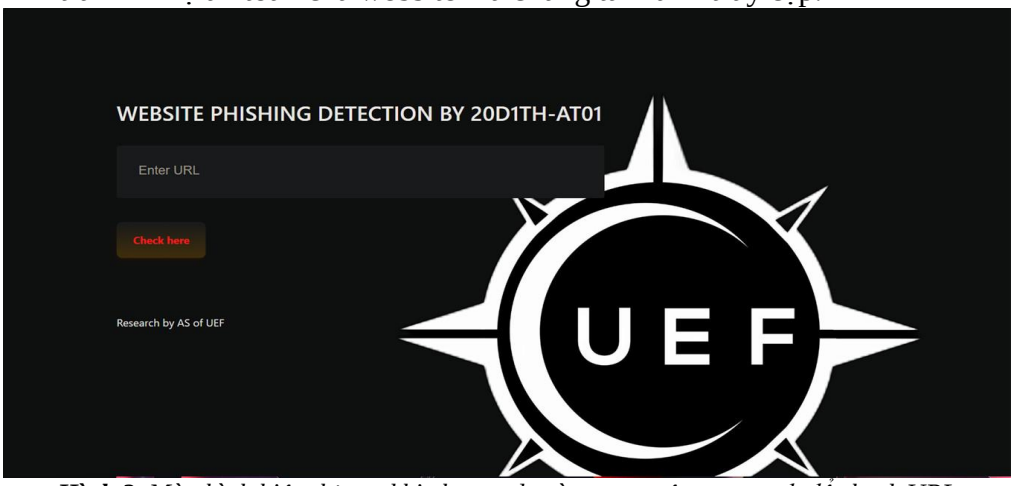
Bảng 2 trình bày kết quả thực nghiệm của các mô hình thuật toán. Để cung cấp cái nhìn tổng quan toàn diện về hiệu suất của mô hình, chúng tôi đánh giá bởi các yếu tố: độ chính xác, tỷ lệ thu hồi và tỷ lệ dương tính giả đều được ghi nhận dưới dạng phần trăm.

Bảng 2: *Kết quả chạy sau khi training bộ dữ liệu với từng thuật toán và kết quả cho thấy thuật toán GBC có độ chính xác cao nhất*

Thuật toán	Accuracy	F1_Score	Recall	Precision
Gradient Boosting Classifier	0.974	0.977	0.994	0.986
Multi-layer Perceptron	0.972	0.975	0.987	0.989
Random Forest	0.966	0.970	0.994	0.988
Support Vector Machine	0.964	0.968	0.980	0.965
Decision Tree	0.959	0.963	0.991	0.993
K-Nearest Neighbors	0.956	0.961	0.991	0.989
Logistic Regression	0.934	0.941	0.943	0.927

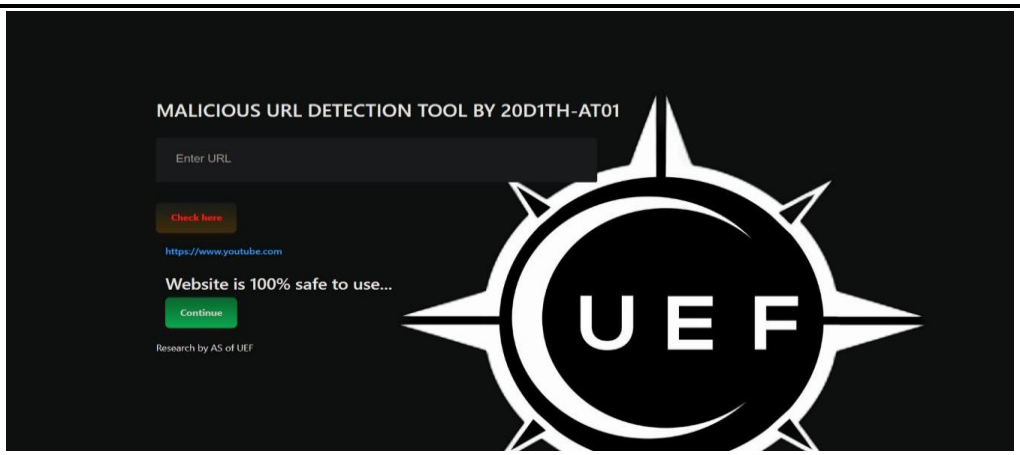
3.4. Công cụ kiểm tra URL giả mạo

Chúng tôi xây dựng công cụ phát hiện tính xác thực của một URL nhằm giúp người dùng có thể kiểm tra mức độ an toàn của website mà chúng ta muốn truy cập.



Hình 3. *Màn hình hiện thị sau khi chạy code và tạo ra một trang web để check URL*

Hình 4. *Kết quả sau khi check URL lừa đảo được lấy từ phishtank*



Hình 5. *Kết quả kiểm tra với trang web YouTube với độ tin cậy là 100%.*

4. Kết Luận

Trong bài báo này, chúng tôi đã trình bày các kỹ thuật phát hiện các URL lừa đảo và đề xuất công cụ kiểm tra độ an toàn của một URL. Trong đó, việc sử dụng thuật toán GBC đã nâng nâng cao việc phát hiện URL lừa đảo với độ chính xác cao hơn với các thuật toán khác. Sau khi huấn luyện dữ liệu thì thuật toán GBC đã cho ra kết quả đáng mong đợi với độ chính xác của thuật toán là 97.4% và F1-score là 97.7%. Tuy nhiên, vẫn còn hạn chế thời gian chờ để kiểm tra URL là lành tính hay độc hại. Trong tương lai, chúng tôi sẽ cải tiến mô hình để cải thiện việc kiểm tra URL nhanh hơn và có thể đạt được việc xét URL lừa đảo hay lành tính một cách chính xác hơn.

Tài liệu tham khảo

[1] https://docs.apwg.org/reports/apwg_trends_report_q2_2023.pdf?

[2] Taeri Kim, Noseong Park, Jiwon Hong, Sang-Wook Kim, *Phishing URL Detection: A Network-based Approach Robust to Evasion* November 2022 Pages 1769–1782 https://doi.org/10.1145/3548606.3560615

[3] Shouq Alnemari, Majid Alshammari, *Detecting Phishing Domains Using Machine Learning*. Appl. Sci. 2023, 13, 4649. https://doi.org/10.3390/app13084649

[4] Tao Feng, Chuan Yue, *Visualizing and Interpreting RNN Models in URL-based Phishing Detection* SACMAT '20: Proceedings of the 25th ACM Symposium on Access Control Models and Technologies June 2020Pages 13–24 https://doi.org/10.1145/3381991.3395602

[5] Sarvagya Agrawal (2021), *Logistic Regression - Supervised Learning Algorithm for Classification* https://www.analyticsvidhya.com/blog/2021/05/logistic-regression-supervised-learning-algorithm-for-classification/

[6] Kirian_Dev Yadav (2023), *K-Nearest Neighbors Algorithm* https://www.linkedin.com/pulse/k-nearest-neighbors-algorithm-kiran-dev-yadav https://apwg.org/trendsreports/

[7] Anshul Sauni (2023),*"Guide on Support Vector Machine (SVM) Algorithm"* https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

[8] IBM, *What is a Decision Tree?*https://www.ibm.com/topics/decision-trees

[9] Simplilearn (2023),*"Random Forest Algorithm"* https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm

[10] Premanand S (2022), *Histogram Boosting Gradient Classifier* https://www.analyticsvidhya.com/blog/2022/01/histogram-boosting-gradient-classifier/

[11] Mayank Bannoula(2023), *An Overview on Multilayer Perceptron (MLP)* https://www.simplilearn.com/tutorials/deep-learning-tutorial/multilayer-perceptron

[12] Eswar Chand (2019), *Phishing website Detector phishing website dataset* https://www.kaggle.com/datasets/eswarchandt/phishing-website-detector