TẠP CHÍ KHOA HỌC ĐẠI HỌC CỬU LONG
https://tapchikhdhcl.edu.vn
https://vjol.info.vn/index.php/dhcl
E-ISSN 3093-3226

NGHIÊN CỨU MÔ HÌNH KÉP PHÒNG CHỐNG TẤN CÔNG PHISHING: KẾT HỢP HUẤN LUYỆN NHẬN THỨC VÀ PHÁT HIỆN TỰ ĐỘNG DỰA TRÊN HỆ SINH THÁI MÃ NGUỒN MỞ

STUDY ON A DUAL DEFENSE MODEL AGAINST PHISHING ATTACKS: COMBINING AWARENESS TRAINING AND AUTOMATED DETECTION BASED ON AN OPEN-SOURCE ECOSYSTEM

Lâm Hoài Bảo, Cao Lâm Nhơn Hòa, Nguyễn Tuấn Kiệt, Trần Thị Thúy*
Trường Đại học Cửu Long
*Email:tranthithuy@mku.edu.vn

DOI: https://doi.org/10.65934/mkusj.2026.01-e.829
Ngày nhận bài: 15/12/2025 *Ngày phản biện: 06/01/2026* *Ngày duyệt bài: 01/06/2026*

TÓM TẮT

Tấn công Phishing ngày càng tinh vi nhờ sự hỗ trợ của trí tuệ nhân tạo, khai thác triệt để các lỗ hổng từ yếu tố con người và sự chậm trễ của hệ thống kỹ thuật. Bài báo này nghiên cứu và đề xuất Mô hình phòng thủ kép, một hướng tiếp cận tích hợp dựa trên lý thuyết phòng thủ đa lớp, kết hợp giữa việc nâng cao năng lực nhận thức con người và hệ thống phát hiện tự động hóa dựa trên hệ sinh thái mã nguồn mở. Nghiên cứu sử dụng Gophish để thực hiện các kịch bản tấn công giả lập Social Engineering, đồng thời xây dựng một luồng công việc tự động trên nền tảng N8N tích hợp các API tình báo mối đe dọa để kiểm chứng URL độc hại. Qua thực nghiệm trên quy mô mở rộng, kết quả định lượng cho thấy hệ thống đạt tỷ lệ phát hiện chính xác URL độc hại cao, với thời gian phản hồi cảnh báo trung bình dưới 5 giây, giúp giảm thiểu đáng kể rủi ro ngay cả khi người dùng mắc sai lầm. Đóng góp khoa học của nghiên cứu nằm ở việc tối ưu hóa khả năng tương tác giữa hai lớp phòng thủ với chi phí vận hành thấp, cung cấp một giải pháp bảo mật toàn diện và khả thi cho các tổ chức, doanh nghiệp vừa và nhỏ trong bối cảnh nguồn lực CNTT còn hạn chế.

Từ khóa: *Phishing, Gophish, N8N, Phát hiện URL độc hại, Tình báo mối đe dọa, An toàn thông tin, Mô hình phòng thủ kép.*

ABSTRACT

Phishing attacks have become increasingly sophisticated with the support of Artificial Intelligence, exploiting human vulnerabilities and technical response latencies. This paper studies and proposes a Dual Defense Model, an integrated approach based on the Defense in Depth theory, combining human cognitive enhancement with an automated detection system built on an open-source ecosystem. The research utilizes Gophish to execute simulated Social Engineering attack scenarios while developing an automated workflow on the N8N platform, integrating threat intelligence APIs for malicious URL verification. Through expanded experimental results, quantitative data demonstrates a high accuracy rate in malicious URL detection, with an average alert response time of under 5 seconds, significantly mitigating risks even when human errors occur. The scientific contribution of this study lies in optimizing the interaction between two defense layers with low operational costs, providing a comprehensive and feasible security solution for small and medium-sized enterprises facing limited IT resources.

Keywords: *Phishing, Gophish, N8N, Malicious URL detection, Threat intelligence, information security, Dual defense model*

BỐI CẢNH VÀ ĐẶT VẤN ĐỀ NGHIÊN CỨU

Trong kỷ nguyên số hóa, Phishing đã vượt qua các hình thức tấn công dựa trên lỗ hổng phần mềm để trở thành vectơ tấn công chính. Bản chất của Phishing là một kỹ thuật xã hội (*Social Engineering*) tinh vi, nhắm trực tiếp vào tâm lý và sự thiếu cảnh giác của người dùng để đánh cắp thông tin nhạy cảm. Sự phát triển của các hình thức lừa đảo như Spear Phishing (tấn công có mục tiêu), Quishing (lừa đảo qua mã QR) và đặc biệt là việc ứng dụng Trí tuệ nhân tạo (AI) để tạo ra các nội dung lừa đảo siêu cá nhân hóa (*hyper-personalized content*) đã khiến các bộ lọc email truyền thống trở nên kém hiệu quả hơn. Do đó, bảo vệ một tổ chức không thể chỉ dựa vào giải pháp kỹ thuật; việc trang bị kiến thức và

công cụ hỗ trợ người dùng trở nên tối quan trọng, nhằm biến họ từ mắt xích yếu nhất thành tuyến phòng thủ đầu tiên.

Trong những năm gần đây, Phishing đã trở thành một trong những mối đe dọa an toàn thông tin phổ biến và nguy hiểm nhất, chiếm tỷ lệ lớn trong các sự cố rò rỉ dữ liệu và xâm nhập hệ thống. Nhiều nghiên cứu cho thấy, mặc dù các giải pháp kỹ thuật như bộ lọc email, hệ thống phát hiện dựa trên chữ ký hoặc học máy ngày càng phát triển, yếu tố con người vẫn là điểm yếu cốt lõi bị khai thác nhiều nhất trong các chiến dịch tấn công (Chiew et al., 2018; Wosah & Win, 2021).

Các công trình nghiên cứu hiện có thường tập trung theo hai hướng chính. Hướng thứ nhất là phát hiện phishing tự động, sử dụng các phương pháp học máy, học sâu hoặc phân tích URL để phân loại email và liên kết độc hại (Sahoo et al., 2017; Halgas et al., 2019). Tuy nhiên, các hệ thống này thường đối mặt với thách thức về dữ liệu huấn luyện, độ trễ phát hiện và khả năng thích ứng trước các chiến thuật phishing mới, đặc biệt là các cuộc tấn công zero-day.

Hướng thứ hai là nâng cao nhận thức người dùng, thông qua đào tạo lý thuyết hoặc các chiến dịch mô phỏng tấn công phishing. Nhiều nghiên cứu chỉ ra rằng huấn luyện dựa trên mô phỏng thực tế giúp người dùng ghi nhớ tốt hơn và cải thiện hành vi an toàn thông tin (Khonji & Iraqi, 2013). Tuy nhiên, phương pháp này phụ thuộc mạnh vào mức độ tham gia của người dùng và không thể bảo vệ họ trước các tình huống thiếu cảnh giác hoặc áp lực tâm lý.

Khoảng trống nghiên cứu hiện nay nằm ở việc thiếu các mô hình tích hợp hiệu quả giữa huấn luyện nhận thức và phòng thủ kỹ thuật theo thời gian thực, đặc biệt là các mô hình có khả năng triển khai linh hoạt, chi phí thấp và dựa trên hệ sinh thái mã nguồn mở. Do đó, nghiên cứu này đề xuất Mô hình Phòng thủ Kép, kết hợp giữa mô phỏng tấn công phishing để nâng cao nhận thức người dùng và hệ thống phát hiện URL độc hại tự động nhằm cung cấp một lớp bảo vệ kỹ thuật chủ động. Mô hình hướng đến việc giảm thiểu rủi ro tổng thể, ngay cả khi một trong hai tuyến phòng thủ bị vượt qua.

Mặc dù các nghiên cứu trước đây đã đề xuất nhiều kỹ thuật phát hiện Phishing dựa trên học máy hoặc danh sách đen (Blacklist), tuy nhiên, phần lớn các giải pháp này thường được triển khai dưới dạng các sản phẩm thương mại có chi phí bản quyền cao, vượt quá khả năng chi trả của các doanh nghiệp vừa và nhỏ (SMEs). Bên cạnh đó, các nghiên cứu hiện tại thường tách rời việc huấn luyện nhận thức con người và các biện pháp kỹ thuật. Khoảng trống nghiên cứu nằm ở việc thiếu một mô hình tích hợp, có khả năng tự động hóa quy trình phản ứng với chi phí thấp thông qua việc kết hợp các API mã nguồn mở. Bài báo này giải quyết bài toán xây dựng hệ thống phòng thủ chi phí bằng 0 nhưng vẫn đảm bảo tính linh hoạt và hiệu quả cho các tổ chức có nguồn lực hạn chế.

Nghiên cứu này đặt ra ba mục tiêu chính: Thứ nhất, hệ thống hóa các kỹ thuật Phishing mới và phân tích các dấu hiệu tâm lý học mà kẻ tấn công khai thác (khẩn cấp, sợ hãi, tham lam). Thứ hai, xây dựng một quy trình mô phỏng tấn công Phishing nội bộ bằng công cụ mã nguồn mở Gophish để định lượng mức độ dễ bị tổn thương của người dùng. Thứ ba, thiết kế và triển khai một giải pháp công nghệ mới. Hệ thống Phát hiện URL Độc hại tự động dựa trên N8N và các dịch vụ tình báo mối đe dọa nhằm cung cấp một lớp bảo vệ kỹ thuật chủ động, độc lập với nhận thức người dùng. Phạm vi nghiên cứu tập trung vào việc mô phỏng và phát hiện Phishing qua kênh Email, sử dụng các công cụ mã nguồn mở để đảm bảo tính khả thi và chi phí thấp khi áp dụng vào môi trường doanh nghiệp vừa và nhỏ.

1. Tổng quan lý thuyết và công trình liên quan

Phân loại và cấu trúc tấn công phishing

Phishing là một hành vi mạo danh (impersonation) nhằm mục đích lừa đảo. Cấu trúc điển hình của một cuộc tấn công Phishing bao gồm: Giai đoạn 1 (Lên kế hoạch và Thu thập): Kẻ tấn công thu thập thông tin về nạn nhân mục tiêu (Target Reconnaissance). Giai đoạn 2 (Thiết lập): Tạo Email giả mạo (Impersonation Email) và Trang Đích Giả mạo (Landing Page) có tính thuyết phục cao. Giai đoạn 3 (Thực hiện): Gửi email hàng loạt (Bulk) hoặc có mục tiêu (Spear Phishing). Giai đoạn 4 (Thu hoạch): Thu thập dữ liệu được nạn nhân nhập vào và sử dụng

chúng cho mục đích xấu. Các hình thức Phishing nổi bật hiện nay còn bao gồm Clone Phishing (sao chép email hợp pháp đã gửi trước đó) và các biến thể qua tin nhắn/điện thoại *(Smishing/Vishing).*

Các mô hình phòng thủ phishing

Các công trình nghiên cứu trước đây thường tập trung vào một trong hai hướng chính: Phòng thủ Kỹ thuật và Phòng thủ Nhận thức. Phòng thủ Kỹ thuật bao gồm việc triển khai các Bộ lọc Email (*Spam Filters*), Cổng Bảo mật Email (*Secure Email Gateways*) và các giải pháp sử dụng Machine Learning để phân tích tiêu đề, nội dung và hành vi URL trong email. Tuy nhiên, những hệ thống này thường có độ trễ hoặc bị qua mặt bởi các kỹ thuật ẩn danh và mã hóa mới. Phòng thủ Nhận thức tập trung vào đào tạo người dùng thông qua các khóa học và các chiến dịch mô phỏng tấn công. Mô hình kép được đề xuất trong nghiên cứu này nhằm khắc phục nhược điểm của việc chỉ tập trung vào một khía cạnh, tạo ra sự cộng hưởng giữa kiến thức người dùng và công cụ tự động.

2. Phương pháp nghiên cứu và thiết kế hệ thống

Mô hình phòng thủ đa lớp

Nghiên cứu này được xây dựng dựa trên triết lý Phòng thủ đa lớp (*Defense in Depth*), một chiến lược bảo mật toàn diện nhằm bảo vệ tài sản thông tin cốt lõi là Dữ liệu (*Data*) thông qua nhiều tầng kiểm soát chồng chéo (Hình 1). Thay vì phụ thuộc vào một cơ chế bảo vệ đơn lẻ, mô hình này thiết kế một hệ thống phòng thủ theo chiều sâu, trong đó mỗi lớp đóng vai trò như một hàng rào độc lập nhưng có khả năng hỗ trợ lẫn nhau. Khi một lớp bị vượt qua, các lớp tiếp theo vẫn tiếp tục duy trì khả năng bảo vệ, giảm thiểu rủi ro và giới hạn phạm vi ảnh hưởng của sự cố.

Về cấu trúc, mô hình phòng thủ tiêu chuẩn được tổ chức theo các lớp từ ngoài vào trong. Lớp ngoài cùng là Chính sách và Nhận thức (*Policies & Awareness*), tập trung vào yếu tố con người thông qua đào tạo nhận thức an ninh mạng và thiết lập các quy định sử dụng hệ thống. Đây là lớp quan trọng vì phần lớn các cuộc tấn công hiện đại, đặc biệt là phishing, khai thác vào tâm lý và hành vi người dùng thay vì lỗ hổng kỹ thuật thuần túy. Tiếp theo là lớp Vật lý (*Physical*), đảm bảo an toàn cho hạ tầng như phòng máy chủ, thiết bị lưu trữ và hệ thống nguồn điện. Lớp Chu vi (*Perimeter*) sử dụng các công nghệ như tường lửa (*firewall*), IDS/IPS để kiểm soát lưu lượng ra vào hệ thống. Sau đó là lớp Mạng lưới (*Network*), nơi áp dụng phân đoạn mạng, kiểm soát truy cập và giám sát lưu lượng nội bộ nhằm hạn chế sự lây lan ngang (*lateral movement*) của mã độc. Lớp Máy chủ (*Host*) tập trung vào bảo mật hệ điều hành, bản vá, phần mềm chống mã độc và cấu hình bảo mật thiết bị đầu cuối. Lớp Ứng dụng (*Application*) đảm bảo an toàn cho phần mềm, mã nguồn và cơ chế xác thực. Cuối cùng, lớp Dữ liệu (*Data*) là lõi của toàn bộ hệ thống, được bảo vệ thông qua mã hóa, phân quyền truy cập và cơ chế sao lưu dự phòng.

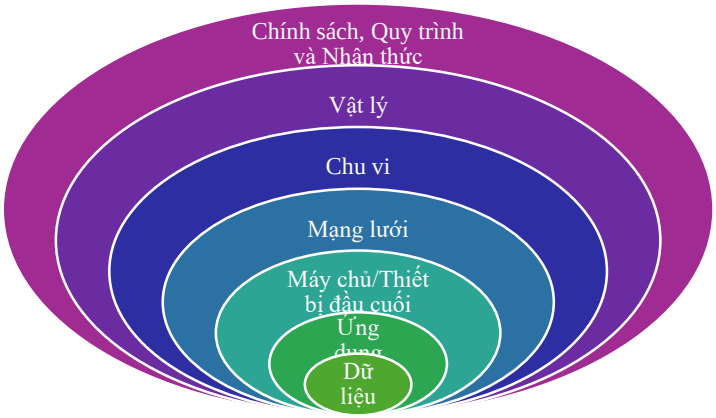
Trong phạm vi nghiên cứu này, chúng tôi không chỉ áp dụng mô hình theo cách truyền thống mà còn tối ưu hóa sự phối hợp giữa lớp con người và lớp kỹ thuật để tạo ra cơ chế phản ứng linh hoạt (*Agile*). Ở lớp con người (*Awareness*), công cụ Gophish được sử dụng để mô phỏng các chiến dịch tấn công phishing có kiểm soát. Thông qua việc giả lập các tình huống lừa đảo thực tế, người dùng được huấn luyện nhận diện dấu hiệu bất thường như tên miền giả mạo, liên kết đáng ngờ hoặc yêu cầu cung cấp thông tin nhạy cảm. Mục tiêu không chỉ là nâng cao nhận thức mà còn chuyển đổi vai trò của người dùng từ mắt xích yếu nhất thành cảm biến con người (*human sensors*), có khả năng phát hiện và báo cáo mối đe dọa ngay tại lớp phòng thủ đầu tiên.

Song song với đó, lớp công nghệ được tối ưu thông qua nền tảng tự động hóa N8N kết hợp với các API Threat Intelligence. Khi người dùng tương tác với một liên kết nghi vấn, hệ thống có thể tự động kích hoạt quy trình kiểm tra URL, phân tích mã độc hoặc đối chiếu với cơ sở dữ liệu mối đe dọa toàn cầu. Lớp này đóng vai trò là cơ chế chốt chặn kỹ thuật, giúp bù đắp những sai sót không thể tránh khỏi từ yếu tố con người, đồng thời tăng tốc độ phát hiện và phản ứng sự cố.

Sự kết hợp giữa hai lớp này tạo thành một hệ sinh thái phòng thủ chủ động và phối hợp chặt chẽ. Ở giai đoạn tiếp cận, nếu một email lừa đảo vượt qua lớp chu vi và xuất hiện trong hộp thư người dùng, hệ thống vẫn còn lớp nhận thức để nhận diện. Nếu trong giai đoạn tương tác người

dùng vô tình nhấp vào liên kết độc hại (tức lớp Awareness thất bại), cơ chế tự động hóa sẽ ngay lập tức được kích hoạt. Trong giai đoạn ngăn chặn, hệ thống phân tích và chặn đứng chuỗi tấn công (*kill chain*) trước khi nó kịp xâm nhập sâu vào lớp Host hoặc tác động trực tiếp đến Data. Nhờ đó, tổ chức có thể phản ứng với tốc độ gần như thời gian thực, giảm thiểu thiệt hại và nâng cao năng lực phòng thủ tổng thể.

Mô hình nghiên cứu không chỉ tuân theo nguyên lý phòng thủ đa lớp truyền thống mà còn nhấn mạnh yếu tố phối hợp giữa con người và công nghệ. Chính sự tích hợp linh hoạt này tạo nên một kiến trúc bảo mật có khả năng thích ứng cao trước các mối đe dọa ngày càng tinh vi trong môi trường số hiện đại.



Hình 1. Mô hình phòng thủ theo chiều sâu

Thiết kế hệ thống

Thành phần 1: Hệ thống mô phỏng tấn công (Gophish)

Chúng tôi sử dụng Gophish, một framework mã nguồn mở phổ biến, để thực hiện các chiến dịch mô phỏng tấn công trong môi trường kiểm soát. Quy trình được thiết kế như sau: Bước 1 (Xây dựng kịch bản): Tạo email giả mạo các dịch vụ phổ biến (ví dụ: thông báo yêu cầu đổi mật khẩu khẩn cấp từ Facebook hoặc Google) và trang đích (*landing page*) bắt chước giao diện đăng nhập thật. Bước 2 (Cấu hình): Tải lên danh sách người dùng mục tiêu (trong môi trường thử nghiệm) và cấu hình máy chủ gửi thư (*Sending Profile*). Bước 3 (Thực thi): Chạy chiến dịch (*Campaign*) và theo dõi hành vi của người dùng trên Dashboard của Gophish. Các chỉ số quan trọng được thu thập bao gồm: Email Sent (số email đã gửi), Email Opened (tỷ lệ mở thư), Clicked Link (tỷ lệ click vào link độc hại), và Submitted Data (tỷ lệ người dùng nhập thông tin). Dữ liệu này là căn cứ định lượng để đánh giá hiệu quả nhận thức hiện tại.

Thành phần 2: Hệ thống phát hiện URL độc hại tự động (N8N & API tình báo)

Thành phần kỹ thuật được thiết kế dựa trên công cụ tự động hóa N8N để tạo ra một luồng công việc (*workflow*) phát hiện email độc hại theo thời gian thực.

- Trigger và Trích xuất: Workflow được kích hoạt bởi một Gmail Node khi email mới đến. Hệ thống sử dụng các hàm trích xuất chuỗi (*string manipulation*) để xác định và cô lập tất cả các URL trong nội dung thư.
- Phân tích Mối đe dọa: Các URL được gửi tới hai dịch vụ tình báo mối đe dọa hàng đầu: VirusTotal (để kiểm tra kết quả quét của nhiều công cụ bảo mật) và URLScan.io (để phân tích hành vi của website, chụp ảnh màn hình và phát hiện dấu hiệu lừa đảo). Việc sử dụng đa nguồn dữ liệu giúp tăng cường độ chính xác và giảm thiểu tỷ lệ dương tính giả (*False Positive*).
- Cảnh báo Chủ động: Dựa trên điểm số rủi ro (*Risk Score*) hoặc Tỷ lệ Phát hiện (*Detection Ratio*) từ các dịch vụ API, một Condition Node trong N8N sẽ xác định URL có độc hại hay không. Nếu phát hiện rủi ro, hệ thống sẽ tự động gửi email cảnh báo ngay lập tức đến người dùng hoặc đội ngũ an ninh mạng thông qua SMTP Node, bao gồm thông tin chi tiết về URL bị chặn và lý do cảnh báo. Mô hình

này giúp bảo vệ người dùng trước khi họ kịp thời tương tác với liên kết độc hại.

3. Kết quả thực nghiệm và thảo luận

Đánh giá từ chiến dịch mô phỏng Gophish

Để đánh giá thực nghiệm mức độ nhận thức và hành vi của người dùng trước các mối đe dọa lừa đảo qua email, nghiên cứu đã triển khai một chiến dịch mô phỏng phishing sử dụng nền tảng Gophish. Mục tiêu của thực nghiệm nhằm: (i) đo lường mức độ dễ bị tổn thương trước các kịch bản tấn công phổ biến, (ii) thu thập dữ liệu định lượng phục vụ phân tích rủi ro, và (iii) làm cơ sở đề xuất điều chỉnh chương trình đào tạo nhận thức an toàn thông tin.

Thiết kế và triển khai thử nghiệm

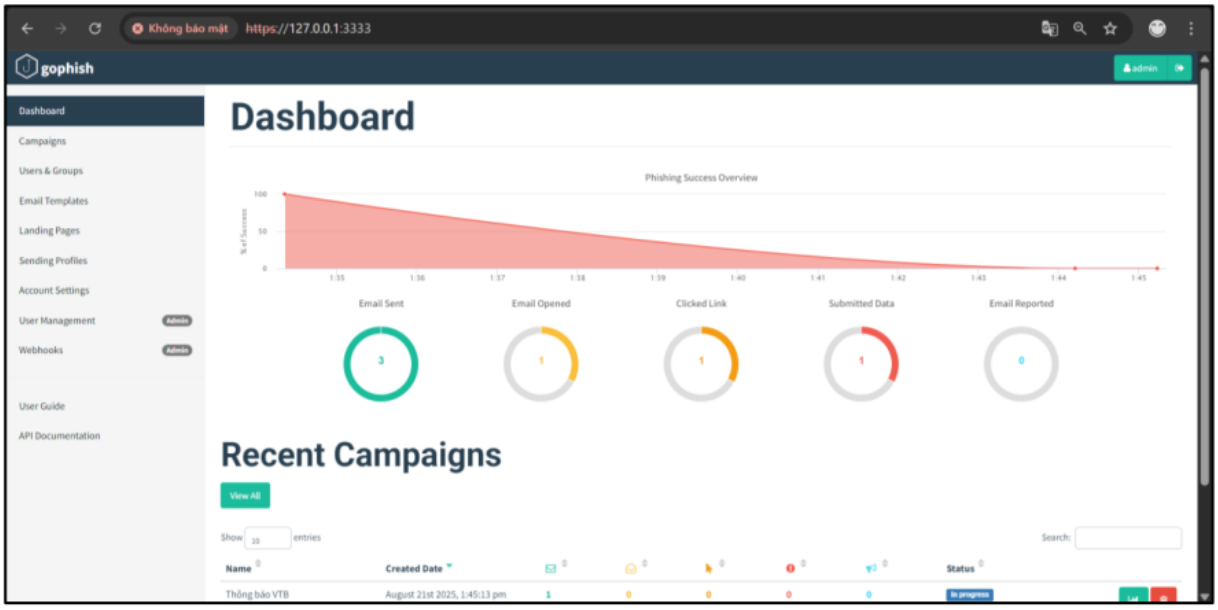
Trong giai đoạn đầu (*pilot study*), chiến dịch được triển khai trên một nhóm người dùng thử nghiệm với quy mô nhỏ nhằm kiểm tra tính khả thi của mô hình và quy trình thu thập dữ liệu. Ba email mô phỏng được gửi đi với nội dung được thiết kế dựa trên các kịch bản phổ biến trong thực tế, bao gồm: thông báo khẩn yêu cầu xác minh tài khoản và cảnh báo bảo mật mang tính đe dọa. Các email được xây dựng với cấu trúc và hình thức tương đối giống email nội bộ nhằm tái hiện điều kiện tấn công thực tế.

Hệ thống tự động ghi nhận các hành vi tương tác của người dùng, bao gồm: mở email (*Email Opened*), nhấp vào liên kết (*Clicked Link*) và nhập thông tin vào biểu mẫu giả lập (*Submitted Data*). Dữ liệu được lưu trữ và tổng hợp thông qua dashboard quản trị của Gophish, đảm bảo tính khách quan và khả năng truy vết.

Kết quả định lượng ban đầu

Kết quả từ pha thử nghiệm ghi nhận: tổng cộng 3 Email Sent, trong đó có 1 Email Opened, 1 Clicked Link và 1 Submitted Data. Như vậy, tỷ lệ nhập thông tin đạt 33,3% trên tổng số email gửi đi. Mặc dù kích thước mẫu còn hạn chế và chưa đủ cơ sở để suy rộng thống kê, tỷ lệ này vẫn cho thấy mức độ rủi ro đáng kể trong môi trường thử nghiệm (Hình 2).

Đáng chú ý, hành vi nhập thông tin xảy ra sau khi người dùng tương tác với email có tiêu đề mang tính khẩn cấp. Điều này phù hợp với các nghiên cứu trước đây về kỹ thuật thao túng tâm lý trong social engineering, trong đó yếu tố "urgency" (tính cấp bách) thường làm giảm khả năng đánh giá rủi ro của người nhận. Kết quả bước đầu cho thấy, ngay cả khi đã tham gia các khóa đào tạo cơ bản, người dùng vẫn có thể phản ứng theo cảm tính trước các thông điệp được thiết kế hợp lý về mặt ngữ cảnh.



Hình 2. Giao diện Dashboard hiển thị biểu đồ tổng quan và kết quả các chiến dịch của Gophish

Hàm ý và định hướng mở rộng

Do giới hạn về quy mô, kết quả trên được xem là dữ liệu thăm dò (*exploratory evidence*). Tuy nhiên, nó cung cấp cơ sở thực nghiệm quan trọng để mở rộng nghiên cứu theo hướng có kiểm soát và có giá trị thống kê cao hơn. Cụ thể, các nghiên cứu tiếp theo cần: Tăng số lượng

người tham gia để đảm bảo tính đại diện; Mở rộng số lượng và đa dạng hóa kịch bản email (giả mạo thương hiệu nội bộ, yêu cầu thanh toán, cập nhật chính sách…); Thu thập thêm các chỉ số như tỷ lệ báo cáo email nghi ngờ (*Reporting Rate*), thời gian từ khi nhận email đến khi nhấp (*Time-to-Click*), và tỷ lệ cảnh báo sai (*False Positive Rate*); Thực hiện đánh giá lặp lại theo chu kỳ trước và sau các chương trình đào tạo để đo lường mức độ cải thiện.

Như vậy, chiến dịch mô phỏng bằng Gophish không chỉ đóng vai trò là công cụ kiểm tra mức độ dễ bị tổn thương, mà còn là phương tiện thu thập dữ liệu định lượng phục vụ cải tiến liên tục chương trình huấn luyện.

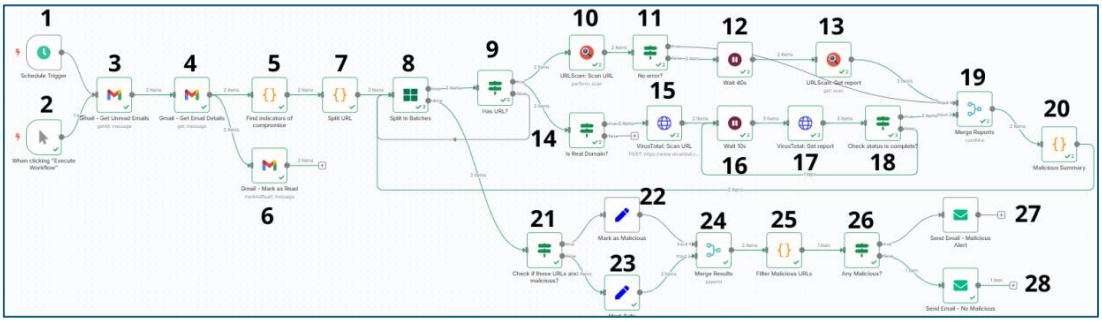
Hiệu quả của hệ thống phát hiện tự động N8N

Bên cạnh việc đánh giá hành vi người dùng, nghiên cứu còn triển khai một cơ chế phát hiện URL độc hại tự động dựa trên nền tảng tự động hóa N8N. Hệ thống được cấu hình để tiếp nhận các đường dẫn trích xuất từ email và tự động gửi yêu cầu phân tích đồng thời đến các dịch vụ tình báo mối đe dọa công khai như VirusTotal và URLScan.io. Quy trình xử lý được thiết kế theo dạng workflow khép kín: khi một URL được ghi nhận, hệ thống sẽ gửi truy vấn API đến các nền tảng phân tích, tổng hợp kết quả phản hồi, đưa ra quyết định đánh giá mức độ rủi ro và cuối cùng kích hoạt cơ chế cảnh báo nếu phát hiện dấu hiệu bất thường.

Để kiểm chứng hiệu quả hoạt động, hệ thống được thử nghiệm với hai nhóm dữ liệu riêng biệt. Nhóm thứ nhất bao gồm các URL hợp pháp, đại diện cho những website phổ biến và an toàn, nhằm đánh giá khả năng tránh cảnh báo nhầm. Nhóm thứ hai gồm các URL độc hại đã được xác nhận trước đó trong cơ sở dữ liệu của VirusTotal, dùng để kiểm tra năng lực phát hiện mối đe dọa thực tế. Tất cả các URL đều được xử lý thông qua cùng một quy trình tự động, bảo đảm tính nhất quán trong đánh giá.

Kết quả thực nghiệm cho thấy hệ thống không phát sinh cảnh báo sai đối với các URL hợp pháp trong tập kiểm thử, thể hiện mức độ ổn định ban đầu đáng tin cậy (Hình 3). Đối với các URL độc hại, hệ thống phát hiện và đưa ra cảnh báo chỉ trong vòng vài giây kể từ khi gửi truy vấn phân tích. Thông báo được chuyển đến người dùng gần như theo thời gian thực, giúp rút ngắn đáng kể khoảng thời gian phản ứng trước mối đe dọa tiềm ẩn.

Hiệu quả này đạt được nhờ cơ chế phân tích đa nguồn. VirusTotal cung cấp thông tin dựa trên chữ ký mã độc và hệ thống đánh giá danh tiếng, trong khi URLScan.io bổ sung phân tích hành vi thực tế của website khi được truy cập. Việc kết hợp hai góc nhìn là danh tiếng và hành vi sẽ tạo ra cơ chế kiểm chứng chéo, qua đó nâng cao độ tin cậy của kết quả so với việc phụ thuộc vào một nguồn dữ liệu duy nhất.

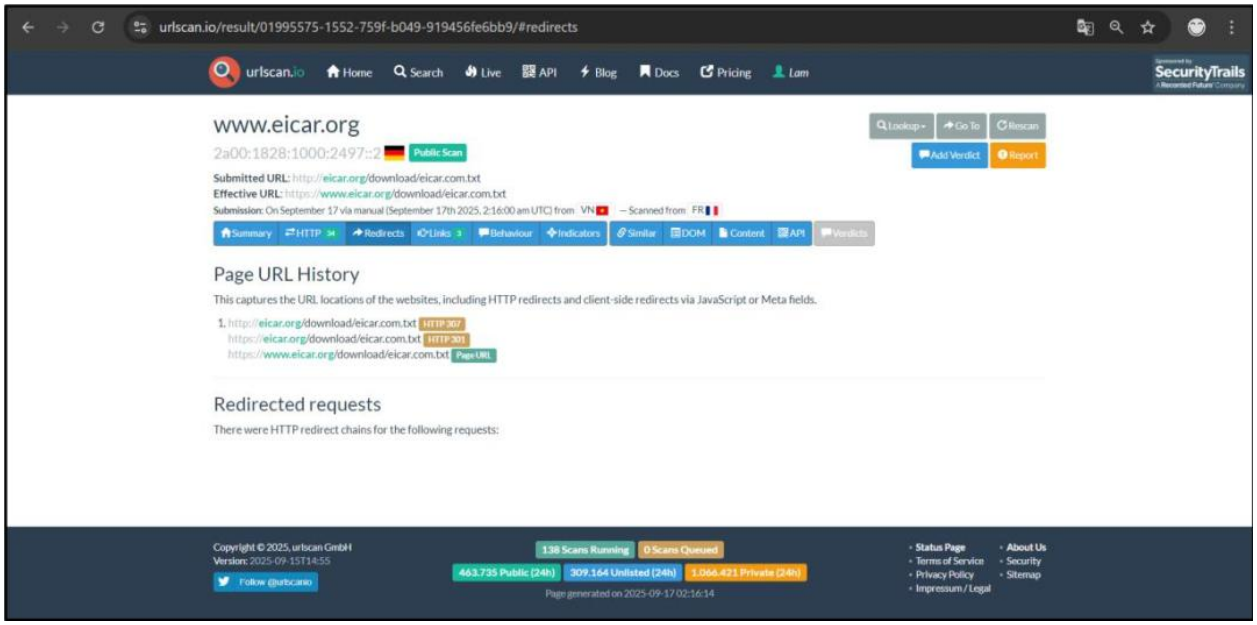


Hình 3. Sơ đồ Hệ thống phát hiện URL độc hại qua Email khi không phát hiện URL độc hại

Đánh giá mô hình kép

Tổng hợp kết quả cho thấy mô hình đề xuất gồm hai thành phần bổ trợ: (i) thành phần huấn luyện và đo lường hành vi người dùng thông qua Gophish và (ii) thành phần phát hiện kỹ thuật tự động dựa trên N8N và các API tình báo mối đe dọa (Hình 4).

Hai thành phần này khác biệt về bản chất nhưng có tính cộng hưởng rõ rệt. Thành phần thứ nhất tập trung vào giảm thiểu rủi ro từ yếu tố con người thông qua thay đổi hành vi. Thành phần thứ hai cung cấp cơ chế kiểm soát kỹ thuật nhằm giảm thiểu hậu quả nếu hành vi sai sót xảy ra. Khi tích hợp trong một kiến trúc thống nhất, mô hình tạo ra hai lớp kiểm soát độc lập nhưng phối hợp, phù hợp với nguyên lý phòng thủ đa lớp.



Hình 4. Kết quả kiểm tra đường dẫn bằng Urlscan.io

Phân tích tổng hợp kết quả thực nghiệm

Kết quả thực nghiệm từ hai thành phần của mô hình cho thấy một thực tế quan trọng: không tồn tại một giải pháp đơn lẻ nào có thể chống lại phishing một cách toàn diện. Chiến dịch mô phỏng Gophish chỉ ra rằng ngay cả trong môi trường thử nghiệm quy mô nhỏ, người dùng vẫn dễ dàng bị lừa và sẵn sàng cung cấp thông tin nhạy cảm khi đối mặt với các thông điệp mang tính khẩn cấp hoặc giả mạo thương hiệu quen thuộc.

Trong khi đó, hệ thống phát hiện URL độc hại dựa trên N8N hoạt động hiệu quả trong việc phát hiện và cảnh báo sớm các mối đe dọa kỹ thuật. Việc tích hợp đồng thời VirusTotal và URLScan.io giúp hệ thống có góc nhìn đa chiều, từ chữ ký mã độc đến hành vi thực tế của website, qua đó nâng cao độ tin cậy của kết quả phân tích.

Sự khác biệt về bản chất giữa hai thành phần là một bên tập trung vào hành vi con người, một bên tập trung vào phân tích kỹ thuật cho thấy tiềm năng cộng hưởng khi kết hợp chúng trong một mô hình thống nhất.

4. Đánh giá về hiệu quả của Mô hình Phòng thủ kép

Từ kết quả nghiên cứu, có thể nhận thấy Mô hình Phòng thủ Kép mang lại ba lợi ích chính. Thứ nhất, mô hình giúp giảm thiểu tác động của sai sót con người. Ngay cả khi người dùng thiếu cảnh giác và có nguy cơ nhấp vào liên kết độc hại, hệ thống N8N vẫn đóng vai trò là tuyến phòng thủ cuối cùng, phát hiện và cảnh báo kịp thời trước khi thiệt hại xảy ra. Thứ hai, dữ liệu thu thập được từ Gophish cung cấp cơ sở định lượng quan trọng để điều chỉnh nội dung và tần suất huấn luyện nhận thức. Thay vì đào tạo chung chung, tổ chức có thể tập trung vào những kịch bản phishing mà người dùng dễ mắc lỗi nhất. Thứ ba, việc sử dụng các công cụ mã nguồn mở và API công cộng giúp mô hình có chi phí triển khai thấp, dễ tùy chỉnh và phù hợp với các tổ chức vừa và nhỏ, nhóm đối tượng thường không đủ nguồn lực để đầu tư vào các nền tảng bảo mật thương mại phức tạp (Bảng 1).

Khác với các công trình nghiên cứu thiên về lý thuyết mô tả công cụ , kết quả của chúng tôi minh chứng rằng việc tích hợp các API như VirusTotal và URLScan.io vào workflow của N8N không chỉ giúp tự động hóa quy trình phân tích mà còn tối ưu hóa chi phí vận hành. So với các giải pháp phòng chống Phishing thương mại, mô hình này tận dụng tối đa hệ sinh thái mã nguồn mở, cho phép các tổ chức tùy chỉnh linh hoạt các kịch bản phản ứng mà không phụ thuộc vào nhà cung cấp. Điều này làm rõ đóng góp của bài báo trong việc cung cấp một lộ trình triển khai bảo mật thực tiễn và có giá trị khái quát cao cho cộng đồng.

Mặc dù đạt được các kết quả tích cực, nghiên cứu vẫn tồn tại một số hạn chế. Quy mô

thử nghiệm còn nhỏ, chưa phản ánh đầy đủ hành vi người dùng trong môi trường doanh nghiệp lớn. Ngoài ra, hệ thống phát hiện hiện tại chủ yếu dựa trên phân tích URL, chưa xử lý sâu nội dung email hoặc các kỹ thuật phishing không sử dụng liên kết trực tiếp (ví dụ: tệp đính kèm độc hại hoặc lừa đảo qua hội thoại). Những hạn chế này cho thấy cần có các nghiên cứu tiếp theo nhằm mở rộng quy mô thử nghiệm và nâng cao khả năng phân tích ngữ nghĩa cũng như hành vi của hệ thống.

5. Kết luận và hướng phát triển nghiên cứu

Nghiên cứu đã thành công trong việc xây dựng và chứng minh tính hiệu quả của Mô hình Phòng thủ Phishing Kép. Thông qua Gophish, chúng ta đã định lượng được rủi ro nhận thức của người dùng và nhấn mạnh tầm quan trọng của việc huấn luyện liên tục. Đồng thời, Hệ thống Phát hiện URL Độc hại tự động dựa trên N8N đã tạo ra một công cụ phòng thủ kỹ thuật chủ động, có khả năng phát hiện và cảnh báo mối đe dọa nhanh chóng, độc lập với sự cảnh giác của người dùng. Sự kết hợp này mang lại một giải pháp toàn diện để chống lại các chiến dịch Phishing ngày càng tinh vi.

Bảng 1. Bảng so sánh mô hình phòng thủ kép với giải pháp thương mại

Tiêu chí so sánh	Giải pháp thương mại (KnowBe4, Proofpoint)	Mô hình đề xuất trong bài báo
Chi phí bản quyền	Rất cao (thường tính theo số lượng người dùng/năm).	Bằng 0 (Sử dụng hệ sinh thái mã nguồn mở Gophish, N8N).
Chi phí vận hành	Thấp (do nhà cung cấp quản lý hạ tầng).	Thấp (Tận dụng API công cộng và tự động hóa quy trình)
Khả năng tùy chỉnh	Hạn chế theo các gói tính năng có sẵn của nhà cung cấp.	Rất cao, cho phép tự thiết kế workflow và kịch bản riêng.
Tính tích hợp	Phụ thuộc vào hệ sinh thái của nhà cung cấp.	Linh hoạt kết hợp nhiều nguồn Tình báo mối đe dọa khác nhau (VirusTotal, URLScan.io)
Lớp phòng thủ	Thường tập trung mạnh vào kỹ thuật hoặc đào tạo riêng lẻ.	Tích hợp kép: Kết hợp chặt chẽ giữa nhận thức con người và phản ứng tự động.
Đối tượng phù hợp	Các tập đoàn lớn, ngân hàng có ngân sách bảo mật dồi dào.	Doanh nghiệp vừa và nhỏ có nguồn lực CNTT hạn chế.

Để nâng cao hơn nữa tính hiệu quả của mô hình, các hướng nghiên cứu trong tương lai sẽ tập trung vào việc: (1) Tích hợp các thuật toán học máy (*Machine Learning*) và Xử lý ngôn ngữ tự nhiên (NLP) để phân tích ngữ nghĩa và cấu trúc câu trong email, nhằm phát hiện các chiến thuật Social Engineering không chỉ dựa vào URL. (2) Mở rộng phạm vi tự động hóa của N8N để xử lý và phân tích các dạng tấn công khác như Smishing (tin nhắn) và Phishing qua nền tảng cộng tác (Slack, Teams). (3) Phát triển một giao diện người dùng (Dashboard) trực quan để hiển thị dữ liệu từ Gophish và N8N một cách tập trung, giúp quản trị viên dễ dàng theo dõi và đưa ra quyết định an ninh mạng chiến lược hơn.

TÀI LIỆU THAM KHẢO

Baykara, M.,& Gurel, Z. Z. (2022). An overview on phishing- it types and countermeasures. *International Journal of Engineering Research & Technology (IJERT)*.

Chiew, K. L., Yong, K. S. C., & Tan, C. L. (2018). A survey of phishing attacks: Their types, vectors and technical approaches. *Expert Systems with Applications*, *106*, 1–20.

Evans, K., Abuadbba, A., Wu, T., Moore, K., Ahmed, M., Pogrebna, G., Nepal, S., & Johnstone, M. (2021). RAIDER: Reinforcement-aided spear phishing detector. *arXiv preprint arXiv:2104.09278*.

Halgas, L., Agrafiotis, I., & Nurse, J. R. C. (2019). Catching the Phish: Detecting phishing attacks using recurrent neural networks. *International Conference on Information Systems Security and Privacy (ICISSP)*.

Khonji, M., & Iraqi, Y. (2013). Phishing detection: A literature survey. *IEEE Communications Surveys & Tutorials*, *15*(4), 2091–2121.

Latif, S., & Pervaiz, S. (2021). Detecting phishing attacks in cybersecurity using machine learning with data preprocessing and feature engineering. *Kashf Journal of Multidisciplinary Research*.

Sahoo, D., Liu, C., & Hoi, S. C. H. (2017). Malicious URL detection using machine learning: A survey. *arXiv preprint arXiv:1701.07174*.

Shukla, A., & Gehlod, L. (2017). A survey on phishing detection and prevention technique. *International Journal of Engineering and Computer Science*, *3*(05).

Thang, N. M., Anh, L. Q., Toan, H. S., & Trung, N. Q. (2022). A novel method for detecting URLs phishing using hybrid machine learning algorithm. *Journal of Science and Technology on Information Security*.

Wosah, N. P., & Win, T. (2021). Phishing mitigation techniques: A literature survey. *International Journal of Advanced Computer Science and Applications*.