Comparative Analysis of Machine Learning Classifiers for Phishing Detection

Mohd Faizal Ab Razak
Faculty of Computing,
College of Computing & Applied Sciences, Universiti Malaysia Pahang
Pahang, Malaysia
faizalrazak@ump.edu.my

Mohd Izham Jaya
Faculty of Computing,
College of Computing & Applied Sciences, Universiti Malaysia Pahang
Pahang, Malaysia
izhamjaya@ump.edu.my

Ferda Ernawan
Faculty of Computing,
College of Computing & Applied Sciences, Universiti Malaysia Pahang
Pahang, Malaysia
ferda@ump.edu.my

Ahmad Firdaus
Faculty of Computing,
College of Computing & Applied Sciences,
Universiti Malaysia Pahang
Pahang, Malaysia
firdausza@ump.edu.my

Fajar Agung Nugroho
Department of Informatics,
Diponegoro University,
Semarang, Indonesia
fajar@lecturer.undip.ac.id

2022 6th International Conference on Informatics and Computational Sciences (ICICoS) | 978-1-6654-6099-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICICoS56336.2022.9930531

***Abstract*— In recent years, communication over the Internet has become the most effective media for leveraging social interactions during the COVID-19 pandemic. Nevertheless, the rapid increase use of digital platforms has led to a significant growth of Phishing Attacks. Phishing attacks are one of the most common security issues in digital worlds that can affects both individual and organization in keeping their confidential information secure. Various modern approaches can be used to target an individual and trick them into leaking their sensitive information, which can later, purposely be used to harm the targeted victim or entire organization depending on the cybercriminal's intent and type of data leaked. This paper evaluates phishing detection by using Naïve Bayes, Simple Logistic, Random Forest, Ada Boost and MLP classifications. This study discusses the comparative analysis on the effectiveness of classification for detecting phishing attacks. The results indicated that the detection system trained with the Random Forest produce higher accuracy of 97.98% than another classifier method.**

Keywords—Phishing Detection, URLs Method, HTML Method, Random Forest Algorithm, Phishing Attack, Online Phishing

I. INTRODUCTION

Nowadays, the world is filled with technologies that benefit most of the people on earth. The most popular one will be the internet. Malicious activity on the internet is increasing at an alarming rate in tandem with the growth of the internet.

This study will focus on one type of malicious activity on the internet which is phishing. Phishing is a scalable act of deception in which a target is impersonated in order to collect information [1]. It is the type of attack usually attached in the email, social media posts/advertisement, or random link popup either direct or indirect to use as bait to access sensitive or personal data.

Based on the statistic released on Statista, the most targeted industries attacked by phishing in the first quarter of 2021 is financial institutions with 24.9% followed by social media with 23.6%. The main goal of the attack mostly is financial gain, and it is one of the reasons why financial institution is on the top of the list. Secondly, social media is the platform where people are sharing information carelessly not knowing that there are people like 'hackers' who pay attentively to their activities and try to benefit from them. The least targeted industry for a phishing attack is cryptocurrency.

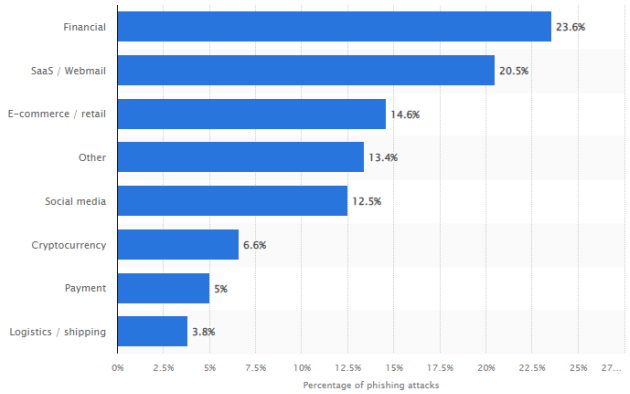

Fig. 1. Percentage of Phishing Attack [2]

There are many methods to prevent phishing attacks use the browser's phishing link, do not click on links or download files from unknown sources, security awareness training, and etc. However, this study will focus on detecting phishing attacks using machine learning. The machine learning algorithm is an automated system code to make an automated decision whether the file contains malicious code or not.

This study presents evaluation of five classifier methods such as Naïve Bayes, Simple Logistic, Random Forest, Ada Boost and MLP classifications for phishing detection. The classifier method which has a highest accuracy are recommended for available solution in security software.

II. LITERATURE REVIEW

In this section, we will discuss five classifier methods for phishing detection systems. The goal of phishing is to infect a user's private system with malware and steal sensitive data from the system without the user's awareness. Machine learning is a sort of artificial intelligence (AI) that can learn without requiring to be programmed explicitly [3]. When exposed to new data, it is capable of forecasting the future and improve decisions. Prediction is usually